Family Wise Error

Learning outcomes:

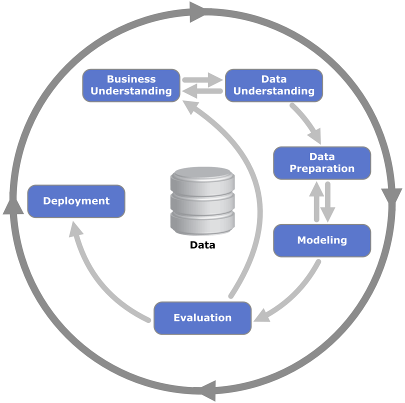


Figure: The CRISP-DM process.

- State the problem with multiple hypothesis tests;
- Define family-wise error;
- Implement multiple correction methods.

Motivation...

- In Lab 6, I asked you to do lots of hypothesis tests;
  - Collect data;
  - Compute test statistic and *p*-value;
  - Reject the null hypothesis if *p*-value $< 0.05 = \alpha$;
- What is α? It is the probability of:

Motivation...

- In Lab 6, I asked you to do lots of hypothesis tests;
  - Collect data;
  - Compute test statistic and *p*-value;
  - Reject the null hypothesis if *p*-value $< 0.05 = \alpha$;
- What is α? It is the probability of:
  - rejecting the null when it is true;
  - a type I error;
  - a false positive;

Motivation...

- In Lab 6, I asked you to do lots of hypothesis tests;
  - Collect data;
  - Compute test statistic and *p*-value;
  - Reject the null hypothesis if *p*-value $< 0.05 = \alpha$;
- What is α? It is the probability of:
  - rejecting the null when it is true;
  - a type I error;
  - a false positive;
- So, if you do lots of hypothesis tests **what is the probability of getting at least one false positive?**

What is the probability of getting at least one false positive?

- If you do one test the probability of no false positive is:

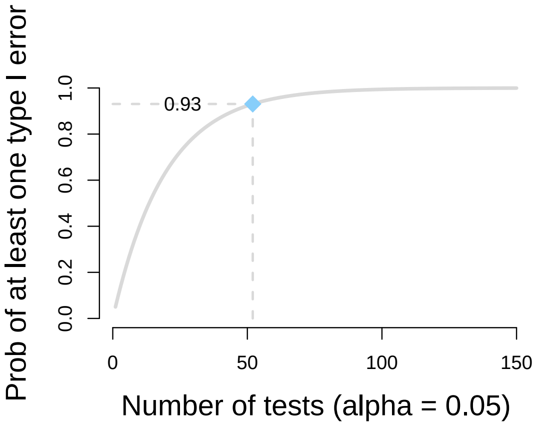
$$1 - \alpha;$$

- If you do k tests the probability of no false positives is:

$$(1 - \alpha)^k;$$

- And so if you do k tests the probability of at least one false positive is:

$$1 - (1 - \alpha)^k.$$



It's not just about hypothesis testing...

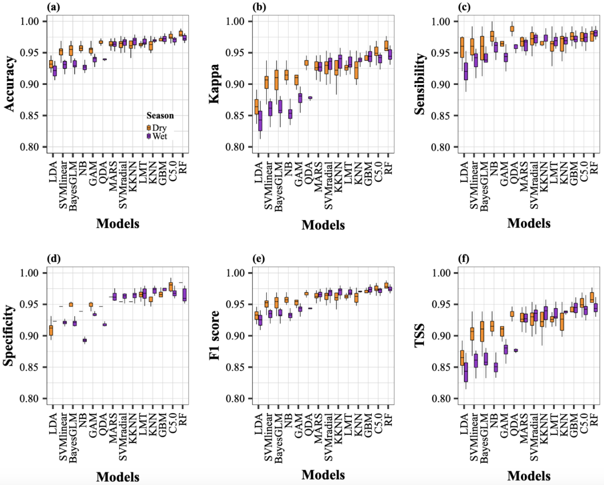


Figure: The authors of "Predicting seasonal movements and distribution of the sperm whale using machine learning algorithms" are doing 168 comparisons in this graphic alone.

Statistical Families

- So the question is:
  - When do we need to worry about this?
  - What groups of tests need to be considered "together"?
  - Which ones do we add up to get k on the previous slide?

Statistical Families

- So the question is:
  - When do we need to worry about this?
  - What groups of tests need to be considered "together"?
  - Which ones do we add up to get k on the previous slide?

- Proposed answer: count all the tests in the same **statistical family** together – a family is:
  - Multiple variables are being tested with no predefined hypothesis (i.e. during EDA);
  - Multiple tests together help support the same research question;
  - Could be tests conducted simultaneously or sequentially over a long period of time;

- For a family of k tests $1 - (1 - \alpha)^k$ is called the **family-wise error rate**!

The Bonferroni Correction

- Suppose you are doing k tests **simultaneously** – reject the null hypothesis if the p-value $\leq \frac{\alpha}{k}$;

- Why? Can show that this makes the family-wise error rate $\leq \alpha$;

$$\text{FWER} = P\left\{\bigcup_{i=1}^{m_0}\left(p_i \leq \frac{\alpha}{m}\right)\right\} \leq \sum_{i=1}^{m_0}\left\{P\left(p_i \leq \frac{\alpha}{m}\right)\right\} = m_0\frac{\alpha}{m} \leq \alpha.$$

- Guarantee: the probability of $\geq$ one type I error with k tests is no more than 0.05.

```
X = ['state'
        ,'longitude (deg)'
        ,'latitude (deg)'
        ,'noaa/temp'
        ,'noaa/altitude'
        ,'male'
        ,'deaths/suicides'
        ,'deaths/homicides'
        ,'bls/2020/unemployed'
        ,'avg_income'
        ,'covid-deaths_total_per_capita'    #constructed
        ,'covid-confirmed_total_per_capita'    #constructed
        ,'covid-vaccination/2021-12-01'
        ,'county_modal_ed'    #constructed
        ,'poverty-rate'
        ,'cost-of-living/living_wage'
        ,'cost-of-living/food_costs'
        ,'cost-of-living/medical_costs'
        ,'cost-of-living/housing_costs'
        ,'cost-of-living/tax_costs'
        ,'health/Average Number of Mentally Unhealthy Days'
        ,'health/% Smokers'
        ,'health/% Adults with Obesity'
        ,'health/% Physically Inactive'
        ,'health/% Long Commute - Drives Alone'
        ,'biggest_industry']    #constructed
```

Figure: Reject null in Lab 6 if?

The Bonferroni Correction

- Suppose you are doing k tests **simultaneously** – reject the null hypothesis if the p-value $\leq \frac{\alpha}{k}$;
- Why? Can show that this makes the family-wise error rate $\leq \alpha$;

$$\text{FWER} = P\left\{\bigcup_{i=1}^{m_0}\left(p_i \leq \frac{\alpha}{m}\right)\right\} \leq \sum_{i=1}^{m_0}\left\{P\left(p_i \leq \frac{\alpha}{m}\right)\right\} = m_0\frac{\alpha}{m} \leq \alpha.$$

- Guarantee: the probability of $\geq$ one type I error with k tests is no more than 0.05.

- What about when we do tests **sequentially**? Suppose:
  - At time 1 you do test 1 and get p-value $= 0.04 < 0.05$, rejecting the null;
  - At time 2 you do test 2 and get p-value $= 0.03 < 0.05$, rejecting the null;
  - Should FAIL to reject the null at both times w/ Bonferroni!

α-spending

- In α-spending:
  - Set a wealth of $W = 0.05$;
  - Require that the sum of the α's for all tests ≤ 0.05;
- For example – for each test halve the remaining budget, $\frac{W}{2^k}$.

	p-val	Reject if $\leq$
test 1	**0.01**	$\frac{W}{2} = 0.025$
test 2	0.06	$\frac{W}{2^2} = 0.0125$
test 3	0.01	$\frac{W}{2^3} = 0.00625$
test 4	**0.003**	$\frac{W}{2^4} = 0.003125$
$\vdots$		$\vdots$
test k		$\frac{W}{2^k}$

α-investing

- In α-investing:
  - Set an initial wealth of W_0 (need not equal 0.05);
  - For test j set: $\alpha_j = \frac{W_{j-1}}{2}$;
  - Update wealth by setting:

$$W_j = \begin{cases} W_{j-1} + 0.05 & \text{if test } j\text{'s } p\text{-value} \leq \alpha_j \\ W_{j-1} - \frac{W_{j-1}}{2 - W_{j-1}} & \text{if test } j\text{'s } p\text{-value} > \alpha_j \end{cases}$$

- So wealth for hypothesis testing grows when you get significant results and decreases when you don't.

	p-val	Reject if $\leq$	W_j
			0.05
test 1	**0.01**	0.025	0.1
test 2	0.06	0.05	0.047
test 3	**0.01**	0.0237	0.097
test 4	**0.003**	0.0487	0.0987
$\vdots$		$\vdots$	

α-debt

- In α-debt:
  - Set an initial $\alpha_0 = 0.05$;
  - For test j set $\alpha_j = \frac{\alpha_0}{j}$;
- So for each new test we apply a Bonferroni correction that treats the family as all previous tests;

	p-val	Reject if $\leq$
		$\alpha_0 = 0.05$
test 1	**0.01**	$\frac{\alpha_0}{1} = 0.05$
test 2	0.06	$\frac{\alpha_0}{2} = 0.025$
test 3	**0.01**	$\frac{\alpha_0}{3} = 0.0167$
test 4	**0.003**	$\frac{\alpha_0}{4} = 0.0125$
$\vdots$		$\vdots$

So how do you choose?