Andrew Gard - equitable.equations@gmail.com

Controlling the family-wise error rate

When conducting multiple statistical tests, the probability of at least one false positive (or type-1 error) increases rapidly, meaning that some random patterns in your data may appear to represent real trends.

When conducting multiple statistical tests, the probability of at least one false positive (or type-1 error) increases rapidly, meaning that some random patterns in your data may appear to represent real trends.

There are many ways of correcting for this, including the **Bonferroni correction** and the **Holm correction**, both of which cap the family-wise error rate (FWER) at a chosen significance level.

When conducting multiple statistical tests, the probability of at least one false positive (or type-1 error) increases rapidly, meaning that some random patterns in your data may appear to represent real trends.

There are many ways of correcting for this, including the **Bonferroni correction** and the **Holm correction**, both of which cap the family-wise error rate (FWER) at a chosen significance level. Neither of these methods make any assumptions about the nature of the statistical tests.

Warning! Data dredging is bad statistical practice even if you correct for multiple testing afterwards.

Warning! Data dredging is bad statistical practice even if you correct for multiple testing afterwards. By combining exploratory and confirmatory analysis, you sacrifice both clarity and statistical power.

Warning! Data dredging is bad statistical practice even if you correct for multiple testing afterwards. By combining exploratory and confirmatory analysis, you sacrifice both clarity and statistical power.

Correcting procedures are most appropriate when ordinary statistical best practices are observed. Hypotheses should be registered before looking at the data which it will be testing (preferably before the data is even collected).

When conducting k independent statistical tests, each with probability α of a false positive, the family-wise error rate is given by

$$FWER = 1 - (1 - \alpha)^k$$

When conducting k independent statistical tests, each with probability α of a false positive, the family-wise error rate is given by

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

For small k, this is approximately $k \times \alpha$. For larger k, it approaches 100%.

The **Bonferroni correction** simply multiplies the p-value of each test by the total number of tests, k.

The **Bonferroni correction** simply multiplies the p-value of each test by the total number of tests, k. This works because the FWER is never greater than k times the significance level of each individual test.

$$FWER \;=\; 1-(1-\alpha)^k \;\leq\; k \times \alpha.$$

The **Bonferroni correction** simply multiplies the p-value of each test by the total number of tests, k. This works because the FWER is never greater than k times the significance level of each individual test.

$$FWER \ = \ 1-(1-\alpha)^k \ \ \leq \ \ k \times \alpha.$$

Multiplying the p-values by k is equivalent to dividing the significance level for each individual test by k.

Example. Suppose 4 tests have p-values of 0.001, 0.005, 0.020, and 0.040.

Example. Suppose 4 tests have p-values of 0.001, 0.005, 0.020, and 0.040.

Using the Bonferroni correction, the respective corrected p-values are 0.004, 0.020, 0.080, and 0.160.

Example. Suppose 4 tests have p-values of 0.001, 0.005, 0.020, and 0.040.

Using the Bonferroni correction, the respective corrected p-values are 0.004, 0.020, 0.080, and 0.160. At significance level $\alpha = .05$, only the first two results would be considered significant.

The Bonferroni correction is an old, well-known, and simple procedure. It doesn't rely on difficult math, and it doesn't make any assumptions about the sorts of tests being performed.

The Bonferroni correction is an old, well-known, and simple procedure. It doesn't rely on difficult math, and it doesn't make any assumptions about the sorts of tests being performed. It's also useful as a back-of-the-envelope mental model when planning a study: k tests at significance level α require a cutoff of α/k for each test.

The Bonferroni correction is an old, well-known, and simple procedure. It doesn't rely on difficult math, and it doesn't make any assumptions about the sorts of tests being performed. It's also useful as a back-of-the-envelope mental model when planning a study: k tests at significance level α require a cutoff of α/k for each test.

On the other hand, the Bonferroni correction is also very conservative, setting a high bar for statistical significance.

The Bonferroni correction is an old, well-known, and simple procedure. It doesn't rely on difficult math, and it doesn't make any assumptions about the sorts of tests being performed. It's also useful as a back-of-the-envelope mental model when planning a study: k tests at significance level α require a cutoff of α/k for each test.

On the other hand, the Bonferroni correction is also very conservative, setting a high bar for statistical significance. This means an increased probability of false negatives (or type-2 errors). That is, some real patterns in your data may be overlooked as potentially just due to random chance.

The **Holm correction** (also known as Holm's step-down procedure or the Bonferroni-Holm correction) is a less conservative approach.

The **Holm correction** (also known as Holm's step-down procedure or the Bonferroni-Holm correction) is a less conservative approach. Like the Bonferroni correction, it controls the family-wise error rate of a collection of tests. However, it often yields smaller adjusted p-values, lowering the probability of type-2 errors.

The **Holm correction** (also known as Holm's step-down procedure or the Bonferroni-Holm correction) is a less conservative approach. Like the Bonferroni correction, it controls the family-wise error rate of a collection of tests. However, it often yields smaller adjusted p-values, lowering the probability of type-2 errors.

The Holm correction is uniformly more powerful than the Bonferroni correction.

To perform the Holm correction on a group of k p-values:

- Arrange the p-values from lowest to highest.

To perform the Holm correction on a group of k p-values:

- Arrange the p-values from lowest to highest.
- Multiply the first p-value by k, the second by $k - 1$, and so on.

To perform the Holm correction on a group of k p-values:

- Arrange the p-values from lowest to highest.
- Multiply the first p-value by k, the second by $k - 1$, and so on.
- Identify the first adjusted p-value that's greater than α. All p-values before this are significant, the others are not.

To perform the Holm correction on a group of k p-values:

- Arrange the p-values from lowest to highest.
- Multiply the first p-value by k, the second by $k - 1$, and so on.
- Identify the first adjusted p-value that's greater than α. All p-values before this are significant, the others are not.
- If no adjusted p-value is greater than α, then all the results are significant.

To perform the Holm correction on a group of k p-values:

- Arrange the p-values from lowest to highest.
- Multiply the first p-value by k, the second by $k - 1$, and so on.
- Identify the first adjusted p-value that's greater than α. All p-values before this are significant, the others are not.
- If no adjusted p-value is greater than α, then all the results are significant.

When reporting a sequence of adjusted p-values like this, it is standard practice to further increase any adjusted p-value that is smaller than the one before it to get a non-decreasing sequence. This is the method used by R's `p.adjust` function.

Example. Suppose we have p-values of $0.005, 0.011, 0.025, 0.035$, and 0.045.

Example. Suppose we have p-values of $0.005, 0.011, 0.025, 0.035$, and 0.045. Individually, these are all significant at $\alpha = .05$.

Example. Suppose we have p-values of $0.005, 0.011, 0.025, 0.035$, and 0.045. Individually, these are all significant at $\alpha = .05$.

- With a Bonferroni correction, the adjusted p-values are 0.025, 0.055, 0.125, 0.175, and 0.225. Only the first result is significant at $\alpha = .05$.

Example. Suppose we have p-values of $0.005, 0.011, 0.025, 0.035$, and 0.045. Individually, these are all significant at $\alpha = .05$.

- With a Bonferroni correction, the adjusted p-values are 0.025, 0.055, 0.125, 0.175, and 0.225. Only the first result is significant at $\alpha = .05$.
- With a Holm correction, the initial adjusted p-values are 0.025, 0.044, 0.075, 0.070, and 0.045. The first two results are significant while the others are not.

Example. Suppose we have p-values of $0.005, 0.011, 0.025, 0.035,$ and 0.045. Individually, these are all significant at $\alpha = .05$.

- With a Bonferroni correction, the adjusted p-values are 0.025, 0.055, 0.125, 0.175, and 0.225. Only the first result is significant at $\alpha = .05$.
- With a Holm correction, the initial adjusted p-values are 0.025, 0.044, 0.075, 0.070, and 0.045. The first two results are significant while the others are not. The adjusted p-values may be reported as 0.025, 0.044, 0.075, 0.075, and 0.075.