Data snooping and the multiple testing fallacy

In 1992 a Swedish study examined whether living near a power line causes adverse health effects. It reported a statistically highly significant increase in childhood leukemia.

Data snooping and the multiple testing fallacy

In 1992 a Swedish study examined whether living near a power line causes adverse health effects. It reported a statistically highly significant increase in childhood leukemia.

A careful follow-up analysis failed to confirm this result.

Why did the study find a statistically significant result?

Data snooping and the multiple testing fallacy

In 1992 a Swedish study examined whether living near a power line causes adverse health effects. It reported a statistically highly significant increase in childhood leukemia.

A careful follow-up analysis failed to confirm this result.

Why did the study find a statistically significant result?

The study looked at 800 different health effects:

There were 800 statistical tests involved.

Multiple comparisons

A statistical test summarizes the evidence for an effect by reporting a p-value:
A smaller p-value means stronger evidence.

Multiple comparisons

A statistical test summarizes the evidence for an effect by reporting a p-value:
A smaller p-value means stronger evidence.

p-value $< 1\%$ $\rightarrow$ test is 'highly significant'

Multiple comparisons

A statistical test summarizes the evidence for an effect by reporting a p-value:
A smaller p-value means stronger evidence.

p-value $< 1\%$ $\rightarrow$ test is 'highly significant'

Interpretation: If there is no effect, then there is only a 1% chance to get such a highly significant result.

Multiple comparisons

A statistical test summarizes the evidence for an effect by reporting a p-value:
A smaller p-value means stronger evidence.

p-value $< 1\%$ $\rightarrow$ test is 'highly significant'

Interpretation: If there is no effect, then there is only a 1% chance to get such a highly significant result.

So if we do 800 tests, then even if there is no effect at all we expect to see
$800 \times 1\% = 8$ highly significant results just by chance!

Multiple comparisons

A statistical test summarizes the evidence for an effect by reporting a p-value:
A smaller p-value means stronger evidence.

p-value $< 1\%$ $\rightarrow$ test is 'highly significant'

Interpretation: If there is no effect, then there is only a 1% chance to get such a highly significant result.

So if we do 800 tests, then even if there is no effect at all we expect to see $800\times$ 1% $= 8$ highly significant results just by chance!

This is called the **multiple testing fallacy** or **look-elsewhere effect**.

Multiple comparisons

A statistical test summarizes the evidence for an effect by reporting a p-value:
A smaller p-value means stronger evidence.

p-value $< 1\%$ $\rightarrow$ test is 'highly significant'

Interpretation: If there is no effect, then there is only a 1% chance to get such a highly significant result.

So if we do 800 tests, then even if there is no effect at all we expect to see $800\times$ $1\% = 8$ highly significant results just by chance!

This is called the **multiple testing fallacy** or **look-elsewhere effect**.

When analyzing large amounts of data it is easy to fall into this trap because there are so many potential relationships to explore, which leads to **data snooping** (=**data dredging**).

Reproducibility and Replicability

Data snooping and other problems have lead to a crisis with regard to **replicability** (getting similar conclusions with different samples, procedures and data analysis methods) and **reproducibility** (getting the same results when using the same data and methods of analysis.)

Reproducibility and Replicability

Data snooping and other problems have lead to a crisis with regard to **replicability** (getting similar conclusions with different samples, procedures and data analysis methods) and **reproducibility** (getting the same results when using the same data and methods of analysis.)

- 'How science goes wrong' in The Economist (10/13/2013)
- 'Why most published research findings are false' by J. Ioannidis (2005)

How can one account for multiple testing?

How can one account for multiple testing?

Bonferroni correction: If there are m tests, multiply the p-values by m.

How can one account for multiple testing?

Bonferroni correction: If there are m tests, multiply the p-values by m.

The Bonferroni correction makes sure that P(any of the m tests rejects in error) $\leq 5\%$.

How can one account for multiple testing?

Bonferroni correction: If there are m tests, multiply the p-values by m.

The Bonferroni correction makes sure that P(any of the m tests rejects in error) $\leq 5\%$.

The Bonferroni correction is often very restrictive: It guards against having even one false positive among the m tests.

How can one account for multiple testing?

Bonferroni correction: If there are m tests, multiply the p-values by m.

The Bonferroni correction makes sure that P(any of the m tests rejects in error) $\leq 5\%$.

The Bonferroni correction is often very restrictive: It guards against having even one false positive among the m tests.

As a consequence the adjusted p-values may not be significant any more even if a noticeable effect is present.

Accounting for multiple testing

Alternatively, we can try to control the **False Discovery Proportion (FDP)**:

$$\text{FDP} = \frac{\text{number of false discoveries}}{\text{total number of discoveries}}$$

where a 'discovery' occurs when a test rejects the null hypothesis.

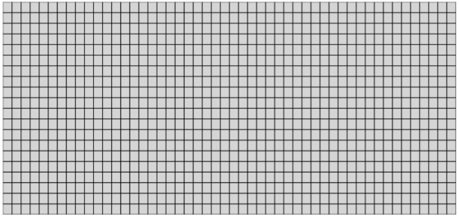
Accounting for multiple testing

Alternatively, we can try to control the **False Discovery Proportion (FDP)**:

$$\text{FDP} = \frac{\text{number of false discoveries}}{\text{total number of discoveries}}$$

where a 'discovery' occurs when a test rejects the null hypothesis.

As an example, we test 1,000 hypotheses.



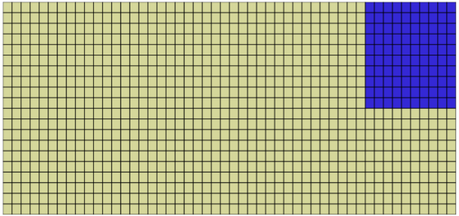
Accounting for multiple testing

Alternatively, we can try to control the **False Discovery Proportion (FDP)**:

$$\text{FDP} = \frac{\text{number of false discoveries}}{\text{total number of discoveries}}$$

where a 'discovery' occurs when a test rejects the null hypothesis.

As an example, we test 1,000 hypotheses.



In 900 cases the null hypothesis is true ("Nothing is going on"), and in 100 cases an alternative hypot hesis is true ("There is an effect: something is going on").

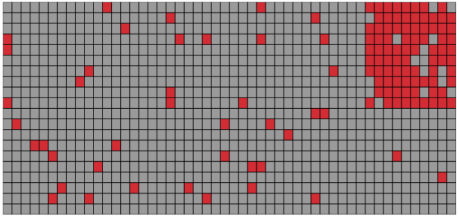
Accounting for multiple testing

Alternatively, we can try to control the **False Discovery Proportion (FDP)**:

$$\text{FDP} = \frac{\text{number of false discoveries}}{\text{total number of discoveries}}$$

where a 'discovery' occurs when a test rejects the null hypothesis.

As an example, we test 1,000 hypotheses.



Doing 1,000 tests results in Discoveries and Non-discoveries.

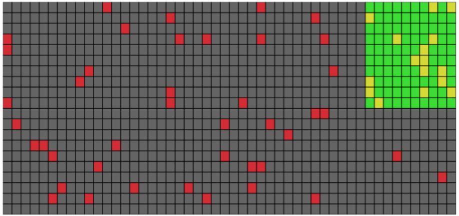
Accounting for multiple testing

Alternatively, we can try to control the **False Discovery Proportion (FDP)**:

$$\text{FDP} \;=\; \frac{\text{number of false discoveries}}{\text{total number of discoveries}}$$

where a 'discovery' occurs when a test rejects the null hypothesis.

As an example, we test 1,000 hypotheses.



We made 80 true discoveries and 41 false discoveries. The false discovery proportion is 41/121=0.34.

Accounting for multiple testing with FDR

False discovery rate (FDR): Controls the expected proportion of discoveries that are false.

Accounting for multiple testing with FDR

False discovery rate (FDR): Controls the expected proportion of discoveries that are false.

Benjamini-Hochberg procedure to control the FDR at level $\alpha = 5\%$ (say):

Accounting for multiple testing with FDR

False discovery rate (FDR): Controls the expected proportion of discoveries that are false.

Benjamini-Hochberg procedure to control the FDR at level $\alpha = 5\%$ (say):

1. Sort the p-values: $p_{(1)} \leq \ldots \leq p_{(m)}$

Accounting for multiple testing with FDR

False discovery rate (FDR): Controls the expected proportion of discoveries that are false.

Benjamini-Hochberg procedure to control the FDR at level $\alpha = 5\%$ (say):

1. Sort the p-values: $p_{(1)} \leq \ldots \leq p_{(m)}$
2. Find the largest k such that $p_{(k)} \leq \frac{k}{m}\alpha$

Accounting for multiple testing with FDR

False discovery rate (FDR): Controls the expected proportion of discoveries that are false.

Benjamini-Hochberg procedure to control the FDR at level $\alpha = 5\%$ (say):

1. Sort the p-values: $p_{(1)} \leq \ldots \leq p_{(m)}$
2. Find the largest k such that $p_{(k)} \leq \frac{k}{m}\alpha$
3. Declare discoveries for all tests i from 1 to k.

Accounting for multiple testing with validation set

Using a validation set: Split the data into a *model-building set* and a *validation set* before the analysis.

Accounting for multiple testing with validation set

Using a validation set: Split the data into a *model-building set* and a *validation set* before the analysis.

You may use data snooping on the model-building set to find something interesting.

Accounting for multiple testing with validation set

Using a validation set: Split the data into a *model-building set* and a *validation set* before the analysis.

You may use data snooping on the model-building set to find something interesting.

Then test this hypothesis on the validation set.

Accounting for multiple testing with validation set

Using a validation set: Split the data into a *model-building set* and a *validation set* before the analysis.

You may use data snooping on the model-building set to find something interesting.

Then test this hypothesis on the validation set.

This approach requires strict discipline: You are not allowed to look at the validation set during the exploratory step!

Mini quiz

1. A medical study examines whether there is a significant correlation between any of 12 lifestyle choices and high blood pressure. It doesn't find any significant correlation, but upon further examination the researchers find a highly significant (p-value <0.5%) correlation between two of the lifestyle choices. This correlation seems not to have been noticed before. Which of the following three statements is an appropriate summary of these findings:

i) The correlation between these two lifestyle choices is highly significant and should be reported as such.

ii) The seemingly significant correlation was found as a consequence of data snooping and therefore the p-value is not valid. The researchers shouldn't report anything.

iii) The seemingly significant correlation was found as a consequence of data snooping and therefore the p-value is not valid. However, this could potentially be a significant new finding. The researchers can report it as such, pointing out that they cannot attach a valid p-value to this finding. It can serve as a hypothesis for a future study with new data, which would then allow for statistically valid conclusions.

2. 1,000 tests were evaluated with the Bonferroni correction. 31 tests had corrected p-values smaller than 5%. Which of the following three statements are an appropriate conclusion:

i) There is a 95% probability that all of these 31 null hypotheses are false.

ii) This is sufficient evidence to reject all of these 31 null hypotheses, because there is only a 5% chance that any of these 31 p-values would be this small if the null hypothesis were true.

iii) If we reject these 31 null hypotheses then we can expect that about 5% of them are rejected in error.

3. 1,000 tests were evaluated with the FDR at the 5% level, which resulted in 31 discoveries. Which of the above statements (i)-(iii) are an appropriate conclusion?