

Andrew Gard - equitable.equations@gmail.com



The dangers of data dredging

Data dredging refers the practice of identifying patterns in a data set through exploratory analysis - either visual or numerical - then selectively testing the observed trends for statistical significance.

Data dredging refers the practice of identifying patterns in a data set through exploratory analysis - either visual or numerical - then selectively testing the observed trends for statistical significance.

Data dredging is bad practice because even random data frequently has patterns in it.

A **type-1 error**, or **false positive**, occurs when a statistical test concludes that a phenomenon is significant even though it's really due to randomness in the sampling procedure.

A **type-1 error**, or **false positive**, occurs when a statistical test concludes that a phenomenon is significant even though it's really due to randomness in the sampling procedure.

The **significance level** α of the test encodes the acceptable risk of this happening. At $\alpha = .05$, a widely-accepted standard, a positive result will be incorrect an average of one time in 20.

When conducting multiple tests in a single study, the probability of a type-1 error grows rapidly.

When conducting multiple tests in a single study, the probability of a type-1 error grows rapidly. The **family-wise error rate (FWER)**, or probability of at least one false positive, is given by

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

where k is the number of tests performed.

When conducting multiple tests in a single study, the probability of a type-1 error grows rapidly. The **family-wise error rate (FWER)**, or probability of at least one false positive, is given by

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

where k is the number of tests performed. For small k , this is approximately $k \times \alpha$. For larger k , it approaches 100%.

When comparing multiple variables pairwise, the situation is even worse.

When comparing multiple variables pairwise, the situation is even worse. With m variables, $\binom{m}{2}$ tests are needed.

When comparing multiple variables pairwise, the situation is even worse. With m variables, $\binom{m}{2}$ tests are needed. At significance level $\alpha = .05$,

m	FWER
3	14.2%
4	26.5%
5	40.1%
6	53.7%
\vdots	\vdots
10	90.1%

With only 6 variables, a false positive is more likely than not. With 10 variables, the probability is over 90%.

The critical r-value for a correlation test with $n = 100$ observations at $\alpha = .05$

Suppose we have $n = 100$ paired observations of two random variables, X and Y . Under the null hypothesis that variables are truly uncorrelated, the test statistic

$$t = \frac{r}{\sqrt{1 - r^2}} \sqrt{n - 2}$$

has a t-distribution with 98 degrees of freedom. At significance level $\alpha = .05$, the critical t-value is

$$t_* = \text{qt}(.975, 98) \approx 1.984.$$

Substituting and solving yields a critical sample correlation of

$$|r_{crit}| \approx .197.$$

While there are statistical techniques for dealing with multiple testing, it's best to separate out exploratory data analysis from formal inference, using different data sets for each.

While there are statistical techniques for dealing with multiple testing, it's best to separate out exploratory data analysis from formal inference, using different data sets for each.

- Use exploratory analysis to generate hypotheses, refine research questions, and determine directions for further work.

While there are statistical techniques for dealing with multiple testing, it's best to separate out exploratory data analysis from formal inference, using different data sets for each.

- Use exploratory analysis to generate hypotheses, refine research questions, and determine directions for further work.
- Use formal inference to check whether specific, previously-stated hypothesis could be plausible or not.

While there are statistical techniques for dealing with multiple testing, it's best to separate out exploratory data analysis from formal inference, using different data sets for each.

- Use exploratory analysis to generate hypotheses, refine research questions, and determine directions for further work.
- Use formal inference to check whether specific, previously-stated hypothesis could be plausible or not.

Note: p-values are sometimes used for model-building, for instance when selecting variables in a regression model. When using them in this way, resist the temptation to also treat them as measuring significance.