**10**

**Avoiding False Discoveries**

Contents

[General 2](#_Toc46067865)

[Significance Testing 3](#_Toc46067866)

[Hypothesis Testing 5](#_Toc46067867)

[Multiple Hypothesis Testing 7](#_Toc46067868)

# General

1. Statistical testing proceeds in a manner analogous to the mathematical proof technique, proof by contradiction, which proves a statement by assuming it is false and then deriving a contradiction. Compare and contrast statistical testing and proof by contradiction.

Statistical testing assumes a result is consistent with the null hypothesis, which is typically the opposite of what we would like to establish, and then rejects this assertion if the p-value of the result is lower than a specified threshold. Unlike proof by contradiction, this conclusion can be erroneous since it is based on probability. If no contradiction can be derived, proof by contradiction does not assume the original statement is true, and analogously, if the p-value of the result is high, then statistical testing does not, or at least, should not, assume that this proves the null hypothesis is true.

2. State whether the following statements are “true” or “false.” No explanation needed.

1. In the context of clustering, the null hypothesis is that there is a cluster structure in the data beyond what might occur at random.

False.

1. Rejecting the null hypothesis doesn’t necessarily imply that we can accept the alternate hypothesis.

True

1. P-value is equal to the probability of the null hypothesis being true, given the observed result.

False.

1. In hypothesis testing, Power + Type I error () = 1

False.

# Significance Testing

1. Which of the following are suitable null hypotheses? If not, explain why.
2. *Comparing two groups* Consider comparing the average blood pressure of a group of subjects, both before and after they are placed on a low salt diet. In this case, the null hypothesis is that a low salt diet does reduce blood pressure.
3. *Classiﬁcation* Assume there are two classes, labeled ‘+’ and ‘–’, where we are most interested in the positive class, e.g., the presence of a disease. H0 is the statement that the class of an object is negative, i.e., that the patient does not have the disease.

(b) is an appropriate null hypothesis since it assumes the opposite of what we would like to establish. By contrast, (a) is not an appropriate null hypothesis since it takes the null hypothesis to be what we would like to establish.

1. State whether the following statements are “true” or “false.” No explanation needed.
2. In the context of clustering, the null hypothesis is that there is a cluster structure in the data beyond what might occur at random.

False.

1. Rejecting the null hypothesis doesn’t necessarily imply that we are accepting the alternate hypothesis.

True.

1. P-value is equal to the probability of the null hypothesis being true, given the observed result.

False.

1. Understanding the p-value
   1. State which of the following are true. You may say that more than one is true. Explanation not required.
      1. A p-value of 0.001 means that the probability the null hypothesis is true is 0.001.
      2. A p-value of 0.001 means that the probability the seeing a test statistic with the observed value or something more extreme is 0.001.
      3. A p-value of 0.001 means that the power of the test is 0.999.

Only (ii) is true. We assume the null hypothesis is true in order to compute the probability of the results null hypothesis. Said another way, the p-value is the probability of the observed test statistic given the null hypothesis, not the probability of the null hypothesis give the data. Also, p-values often are computed even when there is no specific alternative hypothesis. Even if we had a null hypothesis, the power is not computed as 1 – p-value.

* 1. One of your friends claims the p-value she obtained in an experiment absolutely proves the null hypothesis is not true. Could she be correct? Explain very briefly.

Yes, if the p-value is 0.

# Hypothesis Testing

1. Consider the different combinations of effect size and p-value applied to an experiment where we want to determine the efficacy of a new drug.
2. effect size small, p-value small
3. effect size small, p-value large
4. effect size large, p-value small
5. effect size large, p-value large

Whether the effect size is small or large depends on the domain, which for this exercise is medical. For this problem, consider a small p-value to be less than 0.001, while a large p-value is above 0.05. Assume that the sample size is relatively large, e.g., thousands of patients with the condition that the drug hopes to treat.

(a) Which combination(s) would very likely be of interest?

(b) Which combinations(s) would very likely not be of interest?

(c) If the sample size were small, would that change your answers?

1. (iii) would almost certainly be of interest since the p-value indicates that the effect is likely not a statistical fluke and effect size indicates that the effect of the new drug is worthwhile from a medical point of view.
2. (i) and (ii) are likely not interesting from a practical point of view since the effect size is small. (iv) is not interesting since it has a large p-value and may not be trustworthy.
3. Yes. For a small sample, results are much less trustworthy, i.e., it is much more likely that an ineffective drug may show up as effective, and an effective drug may show no or little effect.
4. The following table shows four possible distributions for the number of successes (R) of a binary variable in 10 trials. The distribution of R is binomial and completely determined by the P(X=1), where 1 represents success.



Use the above table for the following calculations.

1. For H0: P(X=1) = 0.6, H1: P(X=1) = 0.7, compute α, β, and power if the critical region is 9 or 10 successes.
2. For H0: P(X=1) = 0.6, H1: P(X=1) = 0.8, compute α, β, and power if the critical region is 9 or 10 successes.
3. For H0: P(X=1) = 0.6, H1: P(X=1) = 0.9, compute α, β, and power if the critical region is 9 or 10 successes.
4. How does power vary as P(X=1) increases?
5. If H0 represents a baseline treatment, and H1 represents a new, hopefully, better treatment, what does your answer in (d) tell us about how easy it is to detect that the new treatment is better in a trial of 10 patients?

(a) 0.0463, 0.8506, 0.1494

(b) 0.0463, 0.6242, 0.3758

(c) 0.0463, 0.2639, 0.7361

(d) Power increases as P(X=1) increases

(e) It is more difficult to detect a treatment that is only somewhat better rather than a lot better, e.g., a treatment with a 70% chance of success vs. a treatment with a 90% chance of success.

# Multiple Hypothesis Testing

1. Consider the following three procedures for multiple hypothesis testing:

1. Naïve Procedure: Every test is evaluated separately without adjusting their significant level
2. Bonferroni Procedure
3. Benjamini-Hochberg (BH) Procedure

Determine which of these procedures will produce the lowest value, medium value, and highest value for two evaluation metrics: Family-wise error rate (FWER) and power. You can answer by filling out the table below. No explanation needed.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Lowest | Medium | Highest |
| FWER |  |  |  |
| Power |  |  |  |

Solution:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Lowest | Medium | Highest |
| FWER | Bonferroni | BH | Naïve |
| Power | Bonferroni | BH | Naive |