MVE055/MSG810 2017 Lecture 9

Petter Mostad

Chalmers

September 25, 2017

Choosing between models

- To get models (e.g., Binomial model, Exponential model, etc. etc.) to use to make predictions, we need to choose between models, and find parameters for the chosen model. We talked last time about finding (estimating) parameters for models; now we look at choosing between models.
- One approach is based on comparing the probability for observing the given data under the two models, when such probabilities can be computed. (In general, the complexities of the models are also compared).
- Using Milton, we focus instead on two classical approaches: Hypothesis testing and significance testing.

Hypothesis testing

1. Two models are established: The *Null hypothesis* H_0 and the *alternative* hypothesis H_1. (H_1 is often what one "wants to show statistically").
2. A *test statistic* T (i.e., a function of a random sample) is established, so that
   - The distribution of the test statistic T can be computed when H_0 is true.
   - The test statistic has one type of values (often small) when H_0 is true and generally another type of values (often large) when H_1 is true.
3. A *rejection region* F is established (generally one or more intervals) and one decides to *reject* H_0 if T is in F while H_0 is *not rejected* if T is not in F.
4. T is computed from observed data, compared with F, and rejected or not based on this.

Properties of hypothesis tests

- Type I and type II errors.
- We assume we can find the distribution of the test statistic T when we assume H_0 is true: Thus we can compute the probability of Type I errors before data is observed. This probability is often denoted α, and called the *significance* of the test. We often choose the rejection region so that $\alpha = 0.05$.
- Similarly, we write β for the probability for Type II errors. This probability cannot always be computed as easily, without further specifying H_1. The *strength* of the test is $1 - \beta$.
- One tries to choose the test statistic maximizing the test strength while the significance is fixed (often at $\alpha = 0.05$).

Example: Tests for expected value of normal distribution

- Assume $X_1, \ldots, X_n$ is a random sample from Normal(μ, σ) with μ, σ unknown. Assume we want to compare $H_0 : \mu = \mu_0$ with $H_1 : \mu \neq \mu_0$ for some fixed μ_0.
- We choose as test statistic

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

  which has a t distribution with $n-1$ degrees of freedom when H_0 is true.
- The rejection region is all T so that $T < -T_0$ or $T > T_0$ for some T_0. To make the significance become α, we must choose

$$T_0 = t_{\alpha/2}$$

  where $t_{\alpha/2}$ is so that $\Pr\left[T > t_{\alpha/2}\right] = \alpha/2$ when T has a t distribution with $n-1$ degrees of freedom.
- Finally we compute our value for T, compare with T_0, and decide to reject H_0 or not based on this.

One-sided and two-sided tests

- The test in the example above is a *two-sided* test. An alternative is a *one-sided* test, where for example $H_0 : \mu \leq \mu_0$ and $H_1 : \mu > \mu_0$.
- The test statistic is the same, but the rejection region becomes all T so that $T > T_0$, where

$$T_0 = t_\alpha$$

  where t_α is so that $\Pr[T < t_\alpha] = \alpha$ when T has a t distribution with $n - 1$ degrees of freedom.
- According to Milton: If H_0 is specified as a *collection of hypotheses* so that the probability for type I error is *maximally* α for every hypothesis, then the hypothesis test still has significance α. Thus we use $\mu = \mu_0$ to compute the significance.
- Correspondingly one can construct a one-sided test with $H_0 : \mu \geq \mu_0$ and $H_1 : \mu < \mu_0$.

Interpreting a hypothesis test

- Correct interpretation: If you use the hypothesis test as a decision rule, then in a (long) series of such decisions, you will reject a proportion α of the correct H_0 hypotheses.
- Note: A hypothesis test does not give you the probability that H_0 or H_1 is true.
- Note: If you reject H_0, it does not mean that H_1 is true.
- Note: If you do not reject H_0, it does not mean it is proven that H_0 is true!
- Note: Wether you reject or not will not only depend on the data, H_0, and H_1, but also on the choice of test statistic.

Significance testing

- Significance testing represents a further development of the ideas of hypothesis testing: Instead of first deciding a significance level α, we compute the value of the test statistic on the data and then the *smallest significance level* α *which would make it possible to reject* H_0 *with this test statistic*.
- This smallest significance level is calle the *p value* of the test.
- Doing the previous hypothesis test with a significance level of 0.05 corresponds to first computing the p-value and then rejecting H_0 if the p value is 0.05 or less.

Example: Significance testing for the expectation of a normal distribution

- Assume $X_1, \ldots, X_n$ is a random sample from Normal(μ, σ) with μ, σ unknown. Assume you would like to compare $H_0 : \mu = \mu_0$ with $H_1 : \mu \neq \mu_0$ for some known μ_0.
- We choose the same test statistic as before:

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

  which has a t distribution with $n-1$ degrees of freedom when H_0 is true.
- We compute the value of T for our data, find $t_{\alpha/2}$ so that $T = -t_{\alpha/2}$ (if $T < 0$) or $T = t_{\alpha/2}$ (if $T > 0$), and use the table for the t-distibution with $n-1$ degrees of freedom to compute α, which becomes the p value.
- In a corresponding way we can compute the p value for one-sided tests.

Interpretation of the p value

- A p-value will tell you something about how "extreme" your data is in the direction of indicating that H_1 is true, compared to random variations in data expected under H_0.
- It is the probability of observing the observed T or "something more extreme" in the direction of H_1, if H_0 is true.
- Some wrong interpretations:
  - The p value does NOT give you the probability that H_0 is true.
  - The p value cannot be directly related to the probability that H_1 is true.
- Remember that the p value may depend on the choice of test statistic, and not only on the data and the hypotheses.

Example

- Assume we are investigating a sequence of independent experiments which each result in success (1) or failure (0). Assume the probability for success is an unknown parameter p. Assume the data from 8 experiments is

$$0, 1, 0, 0, 1, 0, 0, 1$$

  Our null hypothesis for p is $H_0 : p \geq 0.6$, while the alternative hypothesis is $H_1 : p < 0.6$. What is the p value?
- To answer the question we must know which *test statistic* that should be used. Different test statistics give different results.
- Alternative 1: The procedure to obtain the test statistic is to do 8 experiments and let T be the number of successes.
- Alternative 2: The procedure to obtain the test statistic is to continue doing experimentns until 3 successes have been obtained, and let T be the number of experiments done.
- Using alternative 1, we get a p value of 0.174. If we use alternative 2 we get a p value of 0.095. So with a significance level of 0.1, we will reject H_0 using the second test statistic, but not using the first.

Exact computation of the p values in the example

- Alternative 1: If we assume $p = 0.6$ we get $T \sim \text{Binomial}(8, 0.6)$. The possible values for T and their probabilities are given in the table:

0	1	2	3	4	5	6	7	8
0.001	0.008	0.041	0.124	0.232	0.279	0.209	0.090	0.017

The sum of the probabilities for $T = 0, 1, 2$, or 3 is 0.174. This is the probability for the test statistic obtaining the observed value (3) or "something more extreme" in the direction of H_1. Thus it is the p values.

- Alternative 2: If we assume $p = 0.6$ we get $T \sim \text{Neg-Binomial}(3, 0.6)$. The possible values for T and their probabilities are given in the table:

3	4	5	6	7	8	9	10	11
0.216	0.259	0.207	0.138	0.083	0.046	0.025	0.013	0.006

12	13	14	15	16,17,. . .
0.003	0.001	0.001	0.000	totalt 0.000

The sum of the probabilities that $T = 8, 9, 10, \ldots$ is 0.095. This is the probability for the test statistic obtaining the observed value (8) or "something more extreme" in the direction of H_1. Thus it is the p value.

Example: p values in development of new medicines

- Approval of new drugs is a firmly regulated process based on p values.
- Without regulation, it would be possible for drug companies to first perform medical testing and then choose between possible hypotheses, data, and test statistics to report the ones that give the best results for the company.
- To avoid this, companies must submit a detailed description of their experiments *and* how they will obtain their p value *before* the testing starts.
- One consequence is that it is possible for two companies to have performed exactly the same experiments and have obtained exactly the same results, while one gets approval and the other does not, if the first company has made a bet on a better way to compute the p value than the second company.

Usefulness of p values

- According to Milton, p values are "coming into widesprea use" because of their "logical appeal".
- I would instead say that hypothesis testing in general, and p values in particular, are controversial.
- An example: The journal "Basic and Applied Social Psychology" decided on February 2015 to no longer accept papers with methodology depending on p values, as they are controversial.