Hypothesis Testing

Mark Huiskes, LIACS
mark.huiskes@liacs.nl

Estimating sample standard deviation

- Suppose we have a sample x_1, …, x_n
- Average: xbar = (x_1 + … + x_n) / n
- Standard deviation estimate s? Two approaches
  1. s^2 = 1/n * ((x_1 – xbar) + … + (x_n – xbar))
  2. s^2 = 1/(n-1) * ((x_1 – xbar) + … + (x_n – xbar))
- 1. corresponds to the maximum likelihood estimate
- 2. is an unbiased estimate (i.e. E(s) = sig)
- I recommend to use 1.
- Easier way to compute it: s^2 = (sum(x^2) – 1/n * sum(x)) / n. Example: 5 5 6 8 mu=6; (37.5)-> 1.5
- (X – xbar) / (s /sqrt(n)) has a normal density for fixed s. If s is also random, this quantity has a t-density with n or n-1 degrees of freedom, depending on how s was estimated.

Introduction

- Last time: first form of inference: confidence intervals.
- Today hypothesis testing. [Wageningen coffee room: very confusing to students] Let's see…
- Important not to interpret things the wrong way. If you understand the procedure/understand the mathematics, there's no reason to do that, and there's no problem.

Hypothesis Testing Introduction

- Goal: use data to infer if a hypothesis make sense.
- [Start from the sample space]
- Simple hypothesis: completely determines a probability distribution/density on the sample space
- [Draw a picture]
- Somehow define a critical set: if outcome is in that set we reject the hypothesis. [We will see later how we define such sets in practice, but somehow they should capture the unlikely values under the hypothesis. For the other values we say: “Ok, that’s fine, no reason to reject it”
- Example: flipping a coin, counting the number of heads: hypothesis: “this coin is fair” explain the critical region
- Example: hypothesis: “this email is spam”

Some terminology

- Null hypothesis: the hypothesis being tested; [it is assumed true until evidence is found that is strong enough to reject it; the name is a bit of a custom; the negation of the hypothesis is called the alternative hypothesis. This alternative is usually a composite hypothesis. Example H_0: mu = 2; H_a = mu \neq 2]
- [We want to find out if the data gives us reason to reject the null hypothesis.]
- Critical region: subset of sample space of which the outcomes lead to rejection of the specified hypothesis
- [We can now make two types of errors:] Error types:

  type I error (alpha): the null hypothesis is true, but we reject it (‘false alarm’). [Happens when an unlikely event occurs by chance]

  type II error (beta): the null hypothesis is false, but we do not reject it [Can also happen quite easily: e.g. mean is close]
- [Of course, we would like to make the probabilities of these two events as small as possible.]

[Examples]

- Airport security: weapon detection. Null hypothesis: this person is not carrying a weapon.
- → Medicine: Null hypothesis: this medicine does not work.
- → Information Retrieval. Null hypothesis: this document is not relevant to the user
- Spam filter: Null hypothesis: this is an ordinary mail.

- [Talk about these examples in terms of null hypotheses (often a hypothesis that there is no structural effect, but is not necessary), alternative hypotheses; type I errors, type II errors]

Significance level and Power

- [The probability of a type I error is called the significance level of the test]
- Significance level: probability of a type I error (denoted by alpha).
- [Tradeoff between type I and type II errors (explain with spam detector)]
- [The power of a test is related to the type II error]
- Power (denoted by 1-beta) against a SIMPLE alternative hypothesis: the probability that the test correctly rejects the null hypothesis when the alternative hypothesis is true. [DO NOT WIPE OUT; need it later!!!!!]
- [Note beta is the type II error]
- The significance level and the power are probabilities of the same event: that the null hypothesis is rejected. Just computed under different assumptions:
  - Significance level: null hypothesis is true
  - Power: a particular alternative hypothesis is true

  [Explain with the fair coin example]
- A good test has a small significance level [can't help it] (type I error unlikely) and large power (type II error unlikely).

Test Statistic

- The critical/rejection region (CR/RR) is usually formulated using a test statistic.
- Test statistic: quantity computed from the data which has a known distribution/density given the null hypothesis.
- [If the test statistic is outside a certain range, or exceeds a certain threshold, the null hypothesis is rejected.]
- The CR/RR is chosen such that the probability that an outcome is in the rejection region is (at most) the significance level (common values 5%, 1%)
- [Leave p-value/critical value for later].

Basic steps (method 1)

1. Formulate a null hypothesis (and alternative hypotheses)
2. Specify the significance level of the test
3. Choose the procedure to compute a test statistic from the data
4. Determine a RR
5. Collect the data and compute the outcome of the test statistic.
6. Reject the null hypothesis if the test statistic falls in the rejection region.

Hypotheses about parameters (most typical case)

- Null hypothesis: parameter has a certain value, e.g. mu = a.
- Alternative hypotheses mu \neq a
- [Leave one-sided and two-sided tests for later].

Testing for a sample mean (large sample or known s.d.)

- [First an example; then the general procedure]
- Measure IQ of 16 people [say Belgians]. Model: independent trials X_1, …, X16: E(X_i) = mu; V(X_i) = sig^2.
- [We test the hypothesis that their mean IQ = 100]:

  H_0: mu = E(X)= 100
- Suppose we know that sig=12.
- We measure a sample average A_n = (x_1 + .. + x_n) / n = 118.
- Reject or not? [Can the result be reasonably explained by chance?].
- <u>Given that the null hypothesis is true</u>, we know: A_n has a normal density with mean 100 and standard error 12/sqrt(16)=3, so Z = A_n -109/ 3 will have a standard normal density.
- Z is used as the test statistic. [Often a good procedure:] General form: difference with expected value expressed in units of the standard error
- [Let's compute the rejection region, and then backtrack to see what's a good test statistic.]

Rejection Region

- [Before we can compute the rejection region:] Set significance level alpha, e.g. a=0.05.
- [Make a picture of density of Z. When do we reject? When values are too unlikely. So just like last time] Determine values z_alpha such that P(-z_alpha <= Z <= z_alpha) = 1-alpha. RR: Z <= -z_alpha or Z>=z_alpha: z_a = 1.96 (critical value)
- We measured A_n = 118, so z = 109 – 100 / 3 = 3. So: Reject.

General procedure

1. H_0: $E(X) = \mu$
2. Set significance level alpha.
3. $A_n = (X_1 + \ldots + X_n) / n$

   Test statistic: $Z = (A_n - \mu) / \sigma_e$
4. $RR = \{(Z \le -z_\alpha) \text{ or } (Z \ge z_\alpha)\}$ such that $P(RR) = \alpha$. (z_α is the <u>critical value</u>)
5. Compute outcome z of test statistic:

   $z = (a_n - \mu) / \sigma_e$
6. Reject H_0 or not.

Still to do

- Compute/explain p-values (critical values)
- One-sided vs two-sided tests
- Give an example of a power computation
- Apply derived procedure to differences
  - Paired test
  - Difference of means
- If there's time: test coin bias

p-value

- p-value: probability of getting a value of the test statistic as extreme as or more extreme than that observed value, given H_0 is true.
- Explain with example: z=3. Draw a picture:
- p-value = 2 * (0.5 - NA(0,3)) = 2 * (.5 - .4987) = 2 * 0.013 = 0.026
- p-value equal to the significance level at which we would just reject the null hypothesis (i.e. smallest sig level at which we reject)
- Explain this is a different methodology. No need to set anything beforehand. Is often done. “Disadvantage” (only to the very lazy): bit more computation.

One-sided tests

- Explain. Almost always a bad idea, so I won’t show you an example.

Power (example)

- Power of test against the alternative hypothesis that average IQ = 120.
- H_a: mu = 120. Power: probability of rejection
- 1-beta = P(Z <= - z_a or Z >= z_a) (given H_a)
- So: beta = P(-z_a <= X-100/3 <= z_a)= P(-3z_a+100 <= X <= z_a3 +100)= ((-3za+100 -120)/3 <= (X-120)/3 <= 3za+100-120/3)= P(-za-20/3 <= X-120/3 <= za + 20/3)
- Alpha=0.95 means z_a = 1.96, so compute beta = 2 * NA(0, 1.96 + 20/3)=2*NA(0,8.62)=0. So high power against this hypothesis!

Test for Paired observations (paired t-test)

- Assume differences are independent $d1, \ldots, dn$
- $H\_0: E(d\_i) = 0$
- Estimate sample standard deviation of d_i
- $t = (dbar – 0) / s\_d$
- Often applies if you measure the same event with different devices
- Depending situation, normal or t-test.

Hypothesis test about a difference between the means of two large samples

- Model: two independent trials processes, one with mean mu1 and sd sig1, and one with mean mu2 and sd sig2
- [Suppose you want to test if the means of the two samples are the same]
- H0: mu1 = mu2, or: mu1-mu2=0.
- Xbar = Xbar1 – Xbar2 normally distributed under H0
- E(Xbar) = 0; V(Xbar) = s1^2/n1 + s2^2/n2
- Test statistic Z = Xbar – 0 / se (<-sqrt(V(Xbar)))
- If sd's are known are n's large enough, we can use the normal density (else t, but some small complications so not discussed here.)

Example

- 2 brands, see extra paper.

Testing if a coin is fair

- Null hypothesis: the coin is fair. We set a significance level alpha = 0.05. Suppose we measure 40 heads, should we reject the null hypothesis for this significance level?
- phat = N_H / n, N_H ~ Binomial(n, p)
- E(phat) = p
- V(phat) = pq / n
- Standard error sig_e = sqrt(pq / n)
- (phat – p) / sig_e approximately has standard normal density (when testing proportions we always use the normal approximation)
- Critical region: x <= -1.96 or x>1.96
- Test-value: (0.4 – 0.5) / sqrt(0.25/100) = -0.1/0.5 * 10 = -2.
- So, yes we should reject the null hypothesis!
- p-value: P(X>=2 or X<=-2) = 1 – P(-2<= X <= 2) [Warning] = 1- 2NA(0,2) = 1 – 2*.4772 = 0.0456.