Scientific Computing – Statistics

Fabian Sinz
Dept. Neuroethology, University Tübingen
Bernstein Center Tübingen

10/22/2014

Day 3 – study design: choosing n

general theme

1. make an educated guess about the true parameters
2. state how accurate/powerful you want to be
3. select n based on that

estimating a single mean

standard error and α

- Assume you have an estimate s of the standard deviation from the literature.
- The 95% confidence interval is given by

$$\underbrace{|\tilde{\mu} - \mu_0|}_{=:\delta} \geq t_{97.5\%,\nu} \frac{s}{\sqrt{n}}$$

estimating a single mean

standard error and α

- Assume you have an estimate s of the standard deviation from the literature.
- The 95% confidence interval is given by

$$\underbrace{|\tilde{\mu} - \mu_0|}_{=:\delta} \geq t_{97.5\%,\nu} \frac{s}{\sqrt{n}}$$

- How should we choose n to get a confidence interval of a particular size $\pm\delta$?

estimating a single mean

standard error and α

- Assume you have an estimate s of the standard deviation from the literature.
- The 95% confidence interval is given by

$$\underbrace{|\tilde{\mu} - \mu_0|}_{=:\delta} \geq t_{97.5\%,\nu} \frac{s}{\sqrt{n}}$$

- How should we choose n to get a confidence interval of a particular size $\pm\delta$?

  We should set n to be

$$n \geq \left(\frac{t_{97.5\%,\nu} \cdot s}{\delta} \right)^2$$

exercise

choosing n

Example from last lecture: Literature value of thymus gland weights is 34.3g. The estimate of the standard deviation from the literature is $s = 10$g.
The equation for n is

$$n \geq \left(\frac{t_{97.5\%,\nu} \cdot s}{\delta} \right)^2$$

- Assume we want to sacrifice as few animals as possible. We say we are fine with a confidence interval of size $\pm\delta = 5$, how should we choose n?
- What n should we choose for n if we want $\pm\delta = 2$?

Extend your bootstrapping script from yesterday to check that the equation is correct.

How to interrupt for/while loops

- Sometimes you want to stop a for/while loop early.
- The command for that is `break`

Example

```
1 % silly way to find a random number larger than .8
2 for i = 1:2000
3    u = rand();
4    if u >= .8
5      disp('Found it!');
6      break
7    end
8 end
```

winner's curse

Why it is important to estimate n beforehand

Use the thymus gland dataset to repeat the following procedure

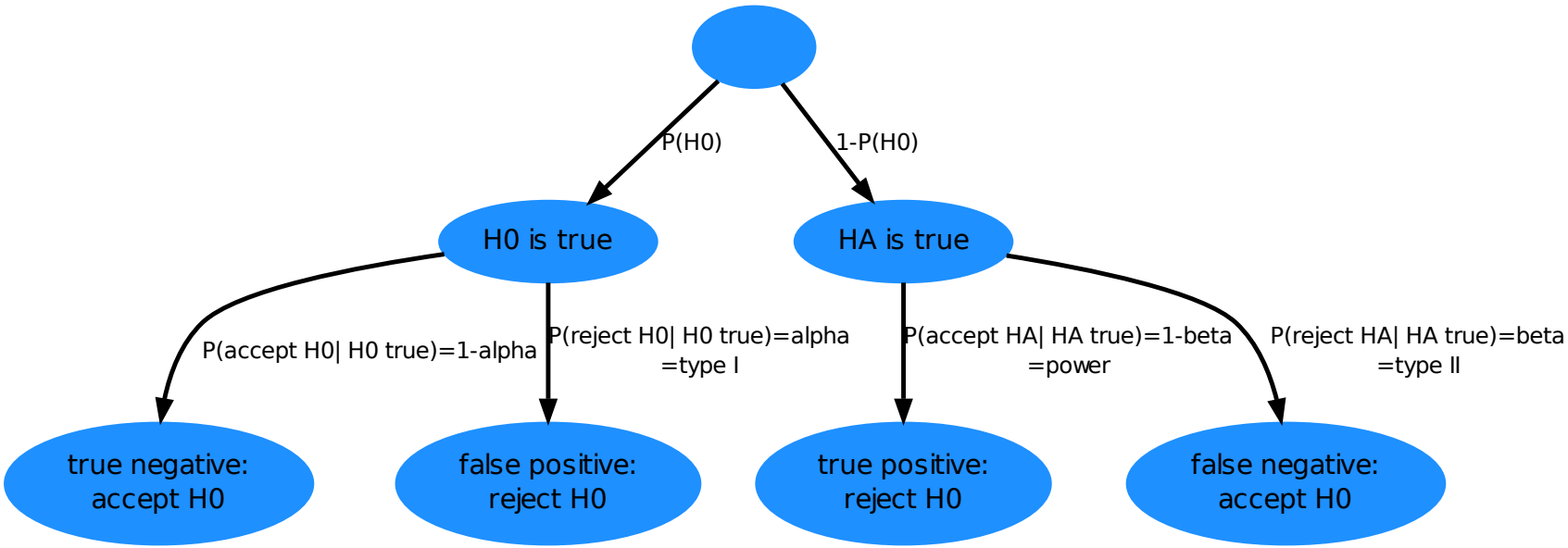
1. Randomly select $n = 10$ numbers from the whole dataset.
2. Perform a one-sample ttest (`ttest`) to test against the mean of 34.3g.
3. If the p-value is smaller than 0.05, stop the loop and print the mean of the 10 datapoints. Also print the mean of the entire thymus gland dataset.
4. Why is it better to use a `for` instead of a `while` loop?
5. What can you observe? Why does that tell you that choosing n is important?

solution

```
load thymusglandweights.dat

n = 10;

x = thymusglandweights;

for i = 1:5000
    idx = randi(length(x), n,1);
    y = x(idx);
    [h,p] = ttest(y, 34.3);

    if h == 1
        disp(['p-value: ', num2str(p)]);
        disp(['mu: ', num2str(mean(y))]);
        disp(['mu total: ', num2str(mean(x))]);
        break
    end
end
```

Day 3 – study design: choosing n

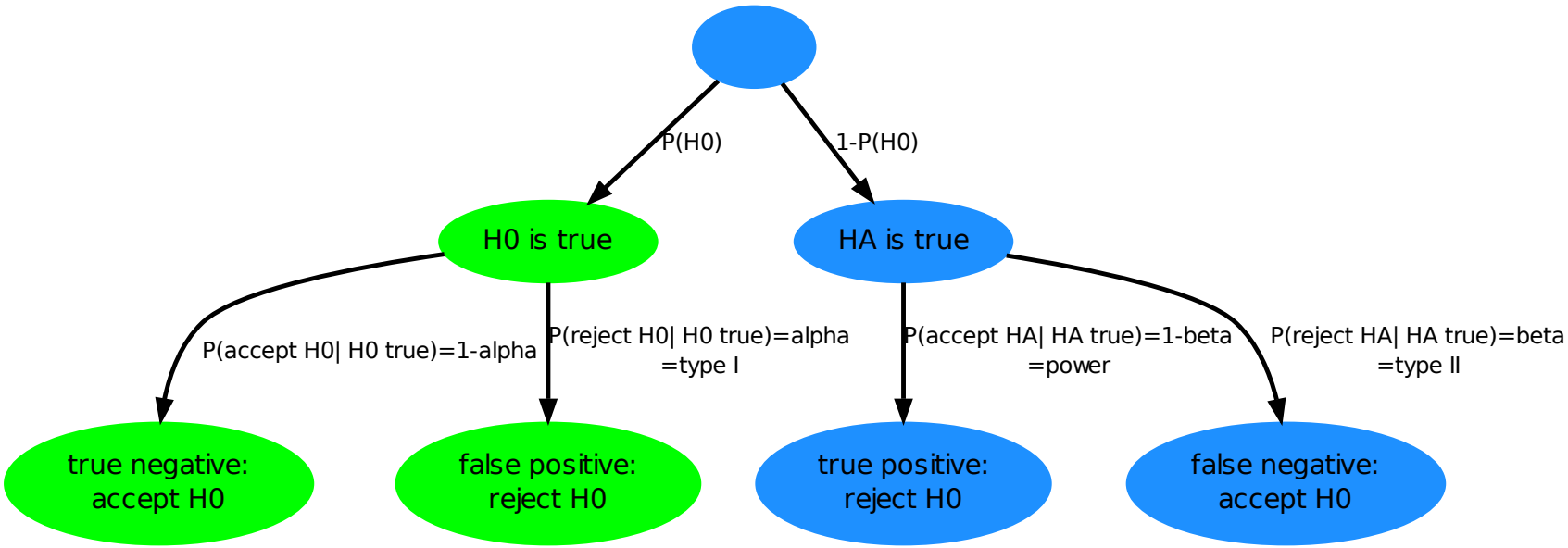
test nomenclature



You want:

- large power
- small type I & II error probability (α and β)

test nomenclature



You want:

- large power
- small type I & II error probability (α and β)

power

estimating power with bootstrapping

- Take the script from yesterday in which we simulated the null distribution of the means.
- Extend it such that it plots the bootstrapped distribution of the means as well (use the same bins for both histograms by using `hist` for computing the histogram and `bar` for plotting).
- Use logical indexing to find all means that correspond to true positives (using the 95% decision boundaries computed yesterday). Estimate the power by computing the fraction of true positive bootstrapped means.
- What is the probability that you get a false negative?
- If you have time, plot the histogram of true positives in a different color.

summary

- Proper study design is important to avoid statistical problems like the winner's curse.
- You should choose a test with high power.
- There are also equations to select n for type I error and power (see book by Zar).

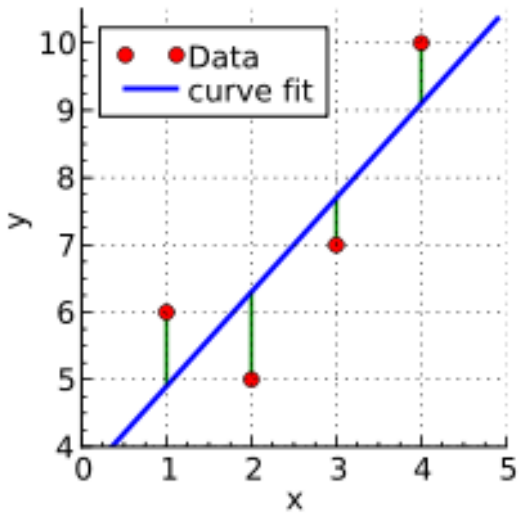
Overview

- minimizing/maximizing a function numerically (optimization) is ubiquitous in science (curve fitting, maximum likelihood, ...)
- today we will look at the basic elements of optimization and apply it to curve fitting
- tomorrow, we will apply it to maximum likelihood

plotting surfaces

```
1 range = linspace(-1,1,20);
2 [X,Y] = meshgrid(range, range);
3
4 surf(X,Y, (X.^2 + Y.^2));
5 colormap('winter');
```

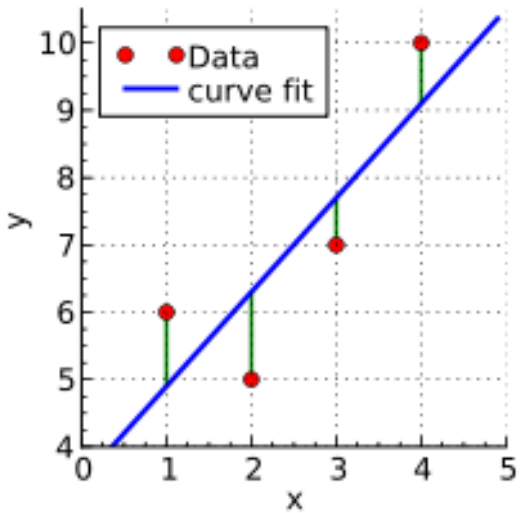
linear least squares



http://en.wikipedia.org/wiki/Linear_least_squares_%28mathematics%29

- The most common curve fitting problem is linear least squares.
- Its goal is to predict a set of output values $y_1, ..., y_n$ from their corresponding input values $x_1, ..., x_n$ with a line $f_{a,b}(x) = ax + b$.
- How is the line chosen?

linear least squares



http://en.wikipedia.org/wiki/Linear_least_squares_%28mathematics%29

- The most common curve fitting problem is linear least squares.
- Its goal is to predict a set of output values $y_1, ..., y_n$ from their corresponding input values $x_1, ..., x_n$ with a line $f_{a,b}(x) = ax + b$.
- How is the line chosen?

  By minimization of the mean squared error

$$g(a, b) = \sum_{i=1}^{n} (y_i - f_{a,b}(x_i))^2$$

error surface

plotting the error surface

Write a function `lserr` that takes 2-dimensional parameter vector (slope and offset), an array of inputs `x`, and an array of corresponding outputs `y`.

That's it.