EMAT30007 Applied Statistics

Lab 4: Hypothesis testing

This lab focusses on the material covered in Lecture 4: testing hypotheses.

1. z-test for a proportion: Exit polls revisited

You are conducting an exit poll for a referendum. You ask $n = 100$ voters at random how they voted. You have 45% of yes in your sample.

- Is there significant evidence at the $\alpha = 1\%$ level that the yes lost the referendum?
- How many voters are needed before 45% of yes becomes significantly smaller than 50% at the $\alpha = 1\%$ level?

Solution

- H_0: the fraction of yes in the population, q, is not smaller than $q_0 = 0.5$.
- Test statistic: $\hat{p} = \bar{x} = 0.45$
- Sampling distribution: Normal $\bar{X} = N(p_0, p_0(1 - p_0)/n)$

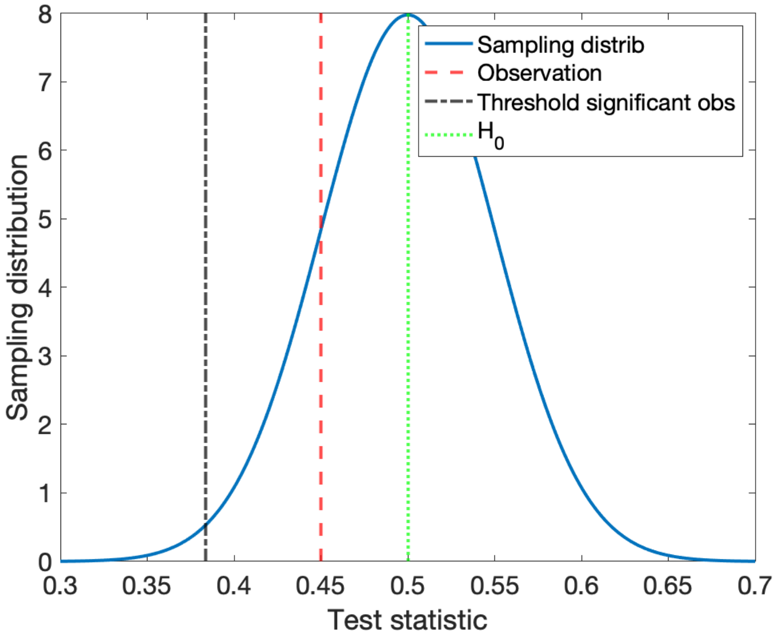
```
% true value
q0 = 0.5;
mu = q0;
n = 100;
sigma = (q0*(1-q0)/n)^0.5;
observation = 0.45;

% left-tailed p-value
pval = normcdf(observation, mu, sigma)
```

```
pval = 0.1587
```

```
% significance level
alpha = 0.01;

% check
[h, p, ci, zval] = ztest(observation, mu, sigma, 'alpha', alpha, 'tail', 'left')
```

```
h = 0
p = 0.1587
ci = 2×1
      -Inf
    0.5663
zval = -1.0000
```

```
% plot
clf;
xx = 0.3:0.001:0.7;
plot(xx, normpdf(xx, mu, sigma), '-', 'LineWidth', 2);
hold on
xline(observation, '--r', 'LineWidth', 2);
hold on
xline(norminv(alpha, mu, sigma), '-.k', 'LineWidth', 2);
hold on
xline(mu, ':g', 'LineWidth', 2);
legend('Sampling distrib', 'Observation', 'Threshold significant obs', 'H_0')
set(gca, 'FontSize', 16.0);
xlabel('Test statistic')
ylabel('Sampling distribution')
```



Given that $p > \alpha$, the fraction of yes is not significanty smaller than 0.5 and we cannot reject the null hypothesis.

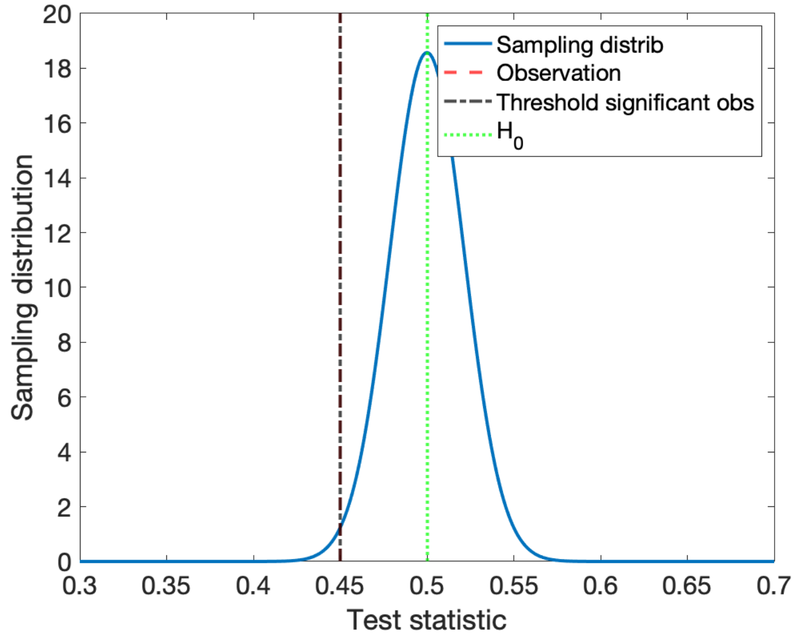
```
% n needed for 0.45 to be 1% significant
n = ceil(((q0*(1-q0))^0.5/(observation - mu) * norminv(alpha))^2)
```

```
n = 542
```

```
sigma = (q0*(1-q0)/n)^0.5;
observation = 0.45;

% left-tailed p-value
pval = normcdf(observation, mu, sigma)
```

```
pval = 0.0100
```

```
% significance level
alpha = 0.01;

% plot
clf;
xx = 0.3:0.001:0.7;
plot(xx, normpdf(xx, mu, sigma), '-', 'LineWidth', 2);
hold on
xline(observation, '--r', 'LineWidth', 2);
hold on
xline(norminv(alpha, mu, sigma), '-.k', 'LineWidth', 2);
hold on
xline(mu, ':g', 'LineWidth', 2);
legend('Sampling distrib', 'Observation', 'Threshold significant obs', 'H_0')
set(gca, 'FontSize', 16.0);
xlabel('Test statistic')
ylabel('Sampling distribution')
```



2. Paired t-test

A streaming media service aims to increase the time its users spend on the website. To this end, a new version of the website has been released, which is expected to be more engaging than the old version. To test this hypothesis, $n = 200$ users have been randomly selected and their website usage has been monitored for two weeks, one week before the website update and one after. For each user, the minutes spent on the old and new versions of the website (m_{before} and m_{after}) have been recorded and saved in the files `m_before.csv` and `m_after.csv`.

- Determine if there has been a significant change in the time spent on the two versions of the website, using a significance level of 0.1%.

To answer the question, assume that the distribution of individual differences $\Delta m = m_{after} - m_{before}$ is Normal with unknown variance and test the hypothesis $\Delta m = 0$ with a t-test.

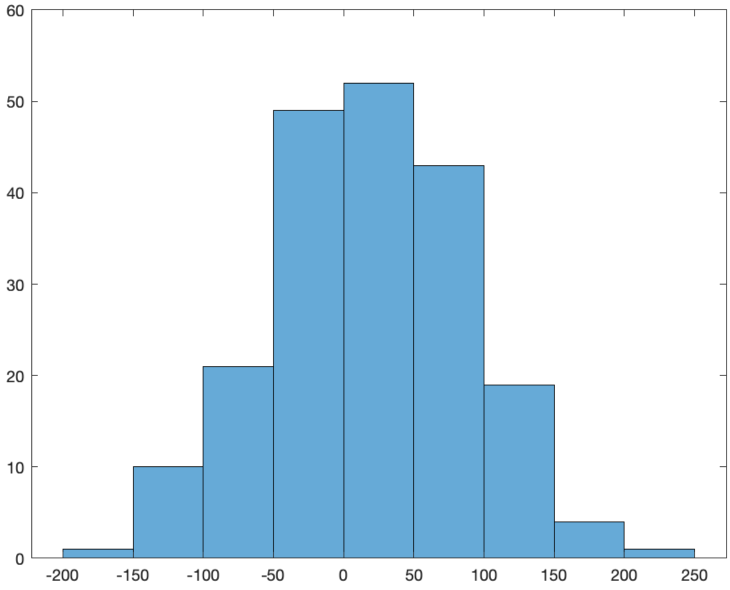
Download the files `m_before.csv` and `m_after.csv` from Blackboard, and place them in Matlab's current folder, which is displayed in the left column, in order to load them using `readmatrix`. Same rows in both files correspond to the same user: for example, the value in the first row of `m_before.csv` is the minutes that user #1 spent on the old version of the website and the value in the first row of `m_after.csv` is the minutes the same user #1 spend on the new version of the website.

Solution

```
m_before = readmatrix("m_before.csv");
m_after = readmatrix("m_after.csv");
Dm = m_after - m_before;
observation = mean(Dm)
```

```
observation = 17.9639
```

```
clf;
histogram(Dm)
```




```
n = length(Dm);
mu = 0;
sigma = std(Dm) / n^0.5;

% test statistic
tstat = (observation - mu)/ sigma
```

```
tstat = 3.6196
```

```
% degrees pf freedom
r = n - 1
```

```
r = 199
```

```

% two-tailed p-value
pval = 2 * (1 - tcdf(tstat, r))
```

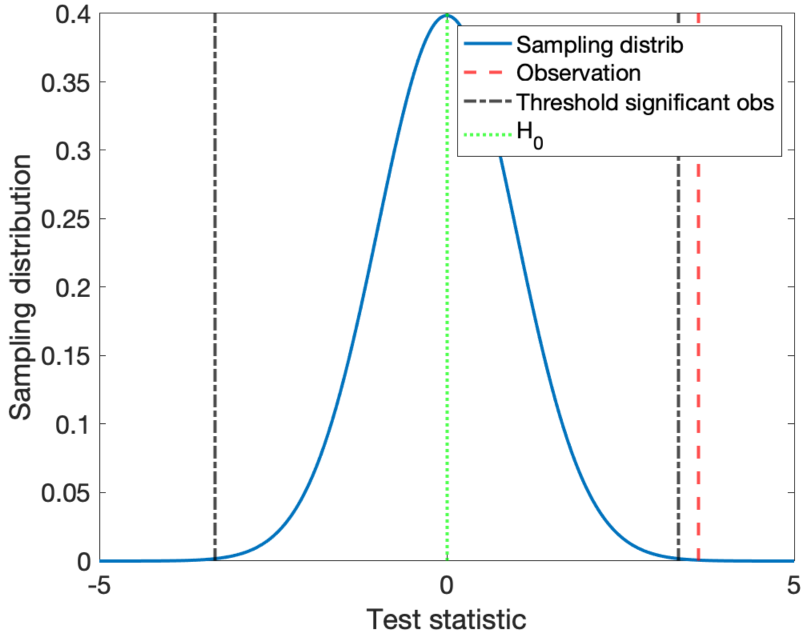
```
pval = 3.7429e-04
```

```
% significance threshold
alpha = 0.001;

% check
[h, p, ci, stats] = ttest(Dm, mu, 'alpha', alpha, 'tail', 'both')
```

```
h = 1
p = 3.7429e-04
ci = 2×1
    1.3871
   34.5407

stats = struct with fields:
    tstat: 3.6196
       df: 199
       sd: 70.1872
```

```
% plot
clf;
xx = -5:0.01:5;
plot(xx, tpdf(xx, r), '-', 'LineWidth', 2);
hold on
xline(tstat, '--r', 'LineWidth', 2);
hold on
xline(tinv(1 -  alpha/2, r), '-.k', 'LineWidth', 2);
hold on
xline(tinv(alpha/2, r), '-.k', 'LineWidth', 2, 'HandleVisibility','off');
hold on
xline(0, ':g', 'LineWidth', 2);
legend('Sampling distrib', 'Observation', 'Threshold significant obs', 'H_0')
set(gca, 'FontSize', 16.0);
```

```
xlabel('Test statistic')
ylabel('Sampling distribution')
```



The probability to obtain a value of the mean difference more extreme than the observed one (two-tailed p-value) is 0.000374, which is smaller than our significance level 0.001. Hence we reject the null hypothesis and conclude that users spent significantly more time on the new version of the website.

3. A/B test

The same streaming media service does another test to verify if the new version of the website is more engaging than the old one. In this second test, two groups of *different* users are randomly selected:

- for each user in the **first group**, the number of minutes spent on the **old version** of the website have been recorded and saved in the file `m_old.csv`
- for each user in the **second group**, the number of minutes spent on the **new version** of the website have been recorded and saved in the file `m_new.csv`.

Determine if there has been a significant change in the time spent on the two versions of the website, using a significance level of 0.1%.

Download the files `m_old.csv` and `m_new.csv` from Blackboard, and place them in Matlab's current folder, which is displayed in the left column, in order to load them using `readmatrix`.

Solution

```
m_new = readmatrix("m_new.csv");
m_old = readmatrix("m_old.csv");
observation = mean(m_new) - mean(m_old)
```

```
observation = 17.9639
```

```
mx = mean(m_new);
sx = std(m_new);
nx = length(m_new);
my = mean(m_old);
sy = std(m_old);
ny = length(m_old);

% degrees of freedom
r = nx + ny - 2;

% test statistic
tstat = (mx - my) / ((sx^2 *(nx-1) + sy^2*(ny-1))/r *(1/nx + 1/nx))^0.5
```

```
tstat = 3.6314
```

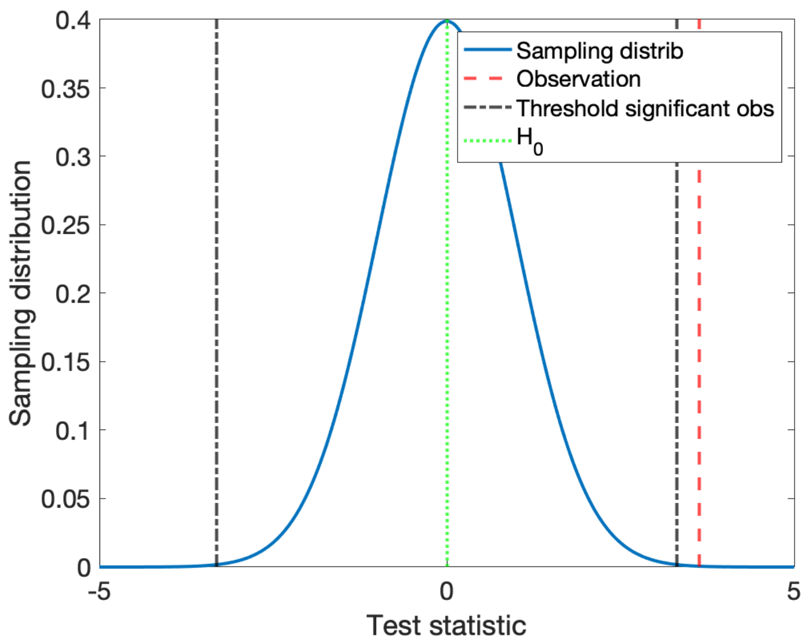
```
% p value
pval = 2 * (1 - tcdf(tstat, r))
```

```
pval = 3.1869e-04
```

```
% significance threshold
alpha = 0.001;

% check
[h, p, ci, stats] = ttest2(m_old, m_new, 'alpha', alpha, 'tail', 'both')
```

```
h = 1
p = 3.1869e-04
ci = 2×1
  -34.3631
   -1.5646

stats = struct with fields:
    tstat: -3.6314
       df: 398
       sd: 49.4678
```

```
% plot
clf;
xx = -5:0.01:5;
plot(xx, tpdf(xx, r), '-', 'LineWidth', 2);
hold on
xline(tstat, '--r', 'LineWidth', 2);
hold on
xline(tinv(1 -  alpha/2, r), '-.k', 'LineWidth', 2);
hold on
xline(tinv(alpha/2, r), '-.k', 'LineWidth', 2, 'HandleVisibility','off');
hold on
xline(0, ':g', 'LineWidth', 2);
legend('Sampling distrib', 'Observation', 'Threshold significant obs', 'H_0')
set(gca, 'FontSize', 16.0);
xlabel('Test statistic')
ylabel('Sampling distribution')
```



The probability to obtain a value of the mean difference more extreme than the observed one (two-tailed p-value) is 0.000318, which is smaller than our significance level 0.001. Hence we reject the null hypothesis and conclude that users spent significantly more time on the new version of the website.

4. Pregnancy test

In pregnancy tests, the presence of the hCG hormone produced by the placenta after implantation is measured: the test result will be "pregnant" if the hCG concentration is significantly higher than the normal level for non-pregnant females. Assuming that

- in non-pregnant females the hCG concentration is Normally distributed with mean 10 and standard deviation 2
- in pregnant females the hCG concentration is Normally distributed with mean 18 and standard deviation 3

and considering the null hypothesis "Not pregnant", calculate:

1. the value of hCG concentration such that the probability to measure a higher value in a non-pregnant female is $\alpha = 0.1$.
2. the probability to commit a type II error (failure to detect pregnancy) if the significance level is $\alpha = 0.1$.
3. the probability to commit a type I error (erroneous detection of pregnancy) if we declare a pregnancy with a hCG concentration of 13 or higher.

Solution

```
mu0 = 10;
std0 = 2;
mu1 = 18;
std1 = 3;

% value of hCG concentration such that the probability to measure
% a higher value in a non-pregnant female is 0.1.
alpha = 0.1;
observation = norminv(1 - alpha, mu0, std0)
```

```
observation = 12.5631
```

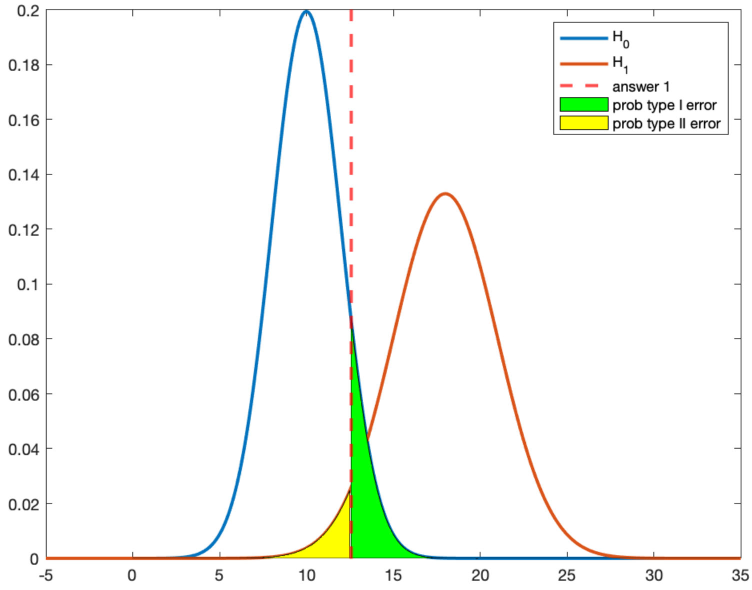
```
% probability to commit a type II error (failure to detect pregnancy)
% if significance level is 0.1
beta = normcdf(observation, mu1, std1)
```

```
beta = 0.0350
```

```
% probability to commit a type I error (erroneous detection of pregnancy)
% if we declare a pregnancy with a hCG concentration of 13 or higher
alpha = 1 - normcdf(13, mu0, std0)
```

```
alpha = 0.0668
```

```
% plot
clf
x = -5:0.1:35;
plot(x, normpdf(x, mu0, std0), '-', 'LineWidth', 2);
hold on
plot(x, normpdf(x, mu1, std1), '-', 'LineWidth', 2);
hold on
xline(observation, '--r', 'LineWidth', 2);
hold on
r = observation:0.1:35;
inBetween = [normpdf(r, mu0, std0), fliplr(zeros(1, length(r)))];
fill([r fliplr(r)], inBetween, 'g');
hold on
r = 0:0.1:observation;
inBetween = [normpdf(r, mu1, std1), fliplr(zeros(1, length(r)))];
fill([r fliplr(r)], inBetween, 'y');
legend('H_0', 'H_1', 'answer 1', 'prob type I error', 'prob type II error')
```



5. Pregnancy test 2 (confusion matrix)

A pharmaceutical company conducted an experiment on a new commercial pregnancy test they have developed.

- A total of 100 random participants took part in the experiment
- 20% of the participants were actually pregnant.
- The probability of a false positive (erroneous detection of pregnancy) is 0.3
- The probability of a false negative (failure to detect pregancy) is 0.1

Compute all the entries of the confusion matrix and calculate:

1. the accuracy of the test, i.e. the fraction of correct predictions
2. the probability to be pregnant if the test result is negative ("not pregnant")
3. the probability of not being pregnant if the test is positive ("pregnant")

Solution

$$\begin{cases} TP+TN+FP+FN & = 100 \\ \dfrac{TP+FN}{TP+TN+FP+FN} & = 0.2 \\ \dfrac{FP}{TN+FP} & = 0.3 \\ \dfrac{FN}{FN+TP} & = 0.1 \end{cases}$$

solving:

$$\begin{cases} TN & = 56 \\ FN & = 2 \\ FP & = 24 \\ TP & = 18 \end{cases}$$

1. Accuracy: $\dfrac{TP+TN}{TP+TN+FP+FN} = 0.74$
2. Fraction of pregnant when test is "not pregnant": $\dfrac{FN}{TN+FN} = 0.03$
3. Fraction of non-pregnant when test is "pregnant": $\dfrac{TP}{TP+FP} = 0.57$