L11 — Hypothesis testing

Statistical hypotheses

- A **statistical hypothesis** is a statement about the distributional properties of data.
- The goal of **hypothesis tests** is to see if data agrees with the statistical hypothesis.
- **Rejection** of a hypothesis indicates that there is sufficient evidence in the data to make the hypothesis unlikely.
- Strictly speaking a **hypothesis test** does not accept a hypothesis, it fails to reject it.

Testing hypotheses

The basis of a hypothesis test consist of

- A **null hypothesis** that we wish to test.
- A **test statistic** \( t(y) \), i.e. a function of the data.
- A **rejection** or **critical region** \( R \).

If the **test statistic** falls in the **rejection** region, then we reject the **null hypothesis**.
Important concepts

**Significance**  The probability (risk) of the test incorrectly rejecting the **null hypothesis**

**Power**  The probability of the test correctly rejecting the **null hypothesis**. Is a function of the true, unknown parameter.

**p-value**  The probability, given the **null hypothesis**, of observing a result at least as extreme as the **test statistic**.

**Type I error**  Incorrectly rejecting the **null hypothesis**.

**Type II error**  Failing to reject the **null hypothesis**.

Statistical hypotheses

We divide hypotheses into

- **Simple**  Specifies a single distribution for the data, e.g. $y \in N(\theta, \sigma^2)$ with $H_0 : \theta = 0$ and $\sigma^2$ known.

- **Composite**  Specifies more than one distribution for the data, e.g. $y \in N(\theta, \sigma^2)$ with $H_0 : \theta = 0$ and $\sigma^2$ unknown; or $y \in N(\theta, \sigma^2)$ with $H_0 : \theta \leq 0$ and $\sigma^2$ known.

Testing simple hypothesis

- In a simple hypothesis the null hypothesis, $H_0$, completely specifies a distribution.
- We construct/define a test statistic, $t(y)$, such that large values of $t(y)$ indicate evidence against $H_0$.
- The p-value of the test is now $p(y) = P(t(Y) \geq t(y)|H_0)$.
- The rejection region is $R = \{y : p(y) \leq \alpha\}$.
- Thus to find the rejection region we need to find the distribution of $t(Y)$, $Y \in H_0$.

Monte Carlo test of a simple hypothesis

**Algorithm**

1. Draw $N$ samples, $y^{(1)}, \ldots, y^{(N)}$, from the distribution specified by $H_0$.
2. Calculate the test statistic, $t^{(i)} = t(y^{(i)})$ for each sample.
3. Estimate the p-value using Monte Carlo integration as

   $$\hat{p}(y) = \frac{1}{N} \sum_{i=1}^{N} 1 \left( t^{(i)} \geq t(y) \right).$$

4. If $\hat{p}(y) \leq \alpha$ reject $H_0$. 
Testing composite hypothesis

For a composite hypothesis $H_0$ defines a set of distributions. Thus, we can’t simply simulate from $H_0$. Three alternatives exist:

- **Pivotal tests** Find a **pivotal** test statistic such that $t(Y)$ has the same distribution for all distributions in $H_0$.
- **Conditional tests** Convert the composite hypothesis to a simple by conditioning on a sufficient statistic.
- **Bootstrap tests** Replacing the composite hypothesis by a simple based on an estimate from data.

Pivotal tests

- In a pivotal test we construct the test statistic so that $t(Y)$ has the same distribution for all distributions specified by $H_0$.
- Example: Assume we have two samples $x_1, \ldots, x_n \in \mathcal{N}(\mu_x, 1)$, $y_1, \ldots, y_m \in \mathcal{N}(\mu_y, 1)$ and want to test if the means are equal:
  $$H_0 : \mu_x = \mu_y, \quad \text{against} \quad H_1 : \mu_x \neq \mu_y$$
  - The test statistic $t(x, y) = |x - y|$ is pivotal as it does not depend on the unknown common mean $\mu$ under $H_0$.
  - Thus, we may use any value of $\mu$ in the simulation step of the Monte-Carlo test.

Rank based tests

- A popular class of pivotal tests are ranked based tests.
- For a sample $y_1, \ldots, y_n$, the rank $r_j$ of $y_j$ is the position of $y_j$ when the sample is sorted with respect to size.
- Examples:
  - **Spearman’s rank correlation** Test if two variables are “correlated”, without making further assumptions about the nature of the relationship between the variables.
  - **Wilcoxon rank-sum test** Test of two samples to see if observations in one of the samples is larger than observations in the other sample.
- Rank based tests assume exchangeability under $H_0$.
- A random variable $Y = (Y_1, \ldots, Y_n)$ is exchangeable if $(Y_{i_1}, \ldots, Y_{i_n})$ has the same distribution for all permutations of the index vector.

Example: pH data

- We have 273 historical and current pH-measurements of 149 lakes in Wisconsin and want to test if the pH-levels have increased.
- We assume that all measurements are independent and that historical measurements have a distribution $F_0$ and that new measurements have a distribution $G_0$.
- We want to test $H_0 : F_0 = G_0$ against $H_1 : F_0 \neq G_0$
Example: pH data

- Under $H_0$, the vector of ranks $r = (r_1, \ldots, r_{273})$ is a draw from the uniform distribution on the set of permutations of the vector $(1, \ldots, 273)$.
- Thus, any statistic based on $r$ is pivotal.
- If the pH-levels had increased, current measurements would have larger ranks, and we choose $t(y)$ as the average difference in ranks.
- This gives a p-value of 0.0131.
- Note: If values in $y$ are not unique we have to account for this by calculating what is called tied ranks.

Example: pH data

- Now use Wilcoxon rank-sum test to test if the observations are equally large in both samples.
- The test statistic is:
  \[
  t = \min \left( \sum r_{1,i} - \frac{n_1(n_1 + 1)}{2}, \sum r_{2,i} - \frac{n_2(n_2 + 1)}{2} \right),
  \]
  where $n_1$ and $n_2$ are the number of observations in the two samples, and $r_{k,i}$ is the rank of samples in the first dataset when the observations are sorted together.
- If all observations in one set is smaller than in the other $t = 0$, thus small values of $t$ indicate evidence against $H_0$: Samples are from the same distribution.
- The p-value is 0.0251, which can be compared with the approximate p-value of 0.0257.

Conditional tests

- In a conditional test we condition on a sufficient statistic to simplify the composite hypothesis.
- A sufficient statistic completely summarizes the information contained in the data about unknown parameters.
- It can be shown that $t(y)$ is a sufficient statistic if and only if the density can be factored as
  \[
  f_\theta(y) = h(y)g_\theta(t(y)).
  \]
- The “trick” is to find a sufficient statistic that does not reduce the distribution of the test statistic to a point mass.
- The p-value of the conditional test is now $p(y) = P(t(Y) \geq t(y)|s(Y) = s(y), H_0)$. 

Example: pH data

```matlab
n1 = length(ph1);
n2 = length(ph2);
% calculate ranks
R = tiedrank([ph1;ph2]);
t = mean(R(n1+1:end))-mean(R(1:n1));
% simulate to get p value
t_perm = zeros(1e5,1);
for i=1:numel(t_perm)
    % create vector of ranks
    R = randperm(n1+n2);
    % calculate U value
    t_perm(i) = mean(R(n1+1:end))-mean(R(1:n1));
end
% p-value
phat = mean(t_perm>=t);
```
Example: Coal mining data

- Recall the coal mining data. From the example in the book we see a clear breakpoint around 1890.
- 1947 the National Coal Board was created to run the nationalized coal mining industry.
- We want to test if this effected the number of accidents.
- \( H_0 \) : \( X_t \) is a Poisson process with intensity \( \lambda \).
- \( H_1 \) : \( X_t \) is Poisson with \( \lambda_1 \) before 1947 and \( \lambda_2 \) after.
- A sufficient statistic for estimating \( \lambda \) is the total number of disasters \( N \).
- Given the number of events in an interval \([T_1, T_2]\), the time of events in a Poisson process are \( U(T_1, T_2) \).
- \( \hat{\lambda}_1 = 1.04 \), \( \hat{\lambda}_2 = 0.50 \) and \( p = 0.054 \).

Permutation tests

- A set of random variables are said to be exchangeable if they have the same distribution for all permutations.
- If the random variables are exchangeable then the ordered variables is a sufficient statistic.
- The distribution of \( y \) given the ordered sample is the uniform distribution on the set of all permutations of \( y \).
- Conditioning on the ordered variables leads to permutation tests.
- Permutation tests can be very efficient in testing an exchangeable null-hypothesis against a non-exchangeable alternative.
- I.e. for testing if two samples differ in some way.

Monte Carlo permutation test

Algorithm

1. Draw \( N \) permutations, \( y^{(1)}, \ldots, y^{(N)} \), of the vector \( y \).
2. Calculate the test statistic, \( t^{(i)} = t(y^{(i)}) \) for each permutation.
3. Estimate the p-value using Monte Carlo integration as

\[
\hat{p}(y) = \frac{1}{N} \sum_{i=1}^{N} 1 \left( t^{(i)} \geq t(y) \right).
\]

4. If \( \hat{p}(y) \leq \alpha \) reject \( H_0 \).
Example: pH data (cont.)

- Assume that the distribution for current data can be written as \( G_0(y) = F_0(y - \theta) \).
- That is, the mean of the current data is the mean of the historical data plus \( \theta \).
- We now want to test \( H_0 : \theta = 0 \) against \( H_1 : \theta > 0 \).
- Under \( H_0 \), all data are iid and thus exchangeable.
- We use the difference in the sample means as a test statistic.
- A permutation test gives \( p = 0.0198 \)

Bootstrap tests

- Remember the relation between hypothesis testing and confidence intervals.
- Knowing how to calculate Bootstrap confidence intervals, we directly have a tool for doing tests on the form \( H_0 : \theta = \theta_0 \) or \( H_0 : \theta > \theta_0 \).
- In general, the idea in a Bootstrap test is to estimate \( P_0 \) from data with \( \hat{P} \).
- Given \( \hat{P} \) we revert to a simple Monte Carlo test, simulating from \( \hat{P} \).
- Note that this is similar to ordinary Bootstrap with the exception that we require \( \hat{P} \) to fulfill the null-hypothesis.
- As an example we compare:
  - \( y \in \mathcal{N}(\theta, \sigma^2) \) with \( H_0 : \theta = 0 \) and \( \sigma^2 \) unknown.

Example: Testing goodness of fit

- Assume that we have some data that we want to test whether it comes from a Gaussian distribution.
- Thus, we want to test \( H_0 : \text{data are iid Gaussian variables} \).
- A natural statistic is the sum of squared differences between the fitted Gaussian cdf and the empirical distribution function:
  \[
  t(y) = \sum_{i=1}^{n} \left( \frac{i}{n} - \Phi \left( \frac{y(i) - \hat{\mu}}{\hat{\sigma}} \right) \right)^2
  \]
- In a Bootstrap test we sample from the fitted Gaussian distribution to estimate the p-value.

Example: Testing goodness of fit

%Estimate parameters:
muhat = mean(y);
sigmahat = std(y);
t = sum(((1:n)/n - normcdf(sort(y),muhat,sigmahat))).^2);
%Bootstrap p-value:
N = 1e5;
t_star = zeros(N,1);
for i=1:N
  %Sample from \( \hat{P} \):
  y_star = sort(muhat + sigmahat*randn(1,n));
  m = mean(y_star);
  s = std(y_star);
  t_star(i) = sum(((1:n)/n - normcdf(y_star,m,s))).^2);
end
phat = mean(t_star>=t)