Course 02403 Introduction to Mathematical Statistics

Lecture 12: Bootstrap and Inference for proportions

DTU Compute
Technical University of Denmark
2800 Lyngby – Denmark

Agenda

1. Bootstrap
   - Parametric bootstrap
   - Non-parametric bootstrapping
2. Inference for proportions
   - Random variable for proportion
   - Hypothesis test for a single proportion
   - Confidence Interval and Hypothesis Test for Two Proportions
   - Hypothesis test for multiple proportions
   - Statistics for contingency tables
3. Summary

Overview

1. Bootstrap
   - Parametric bootstrap
   - Non-parametric bootstrapping
2. Inference for proportions
   - Random variable for proportion
   - Hypothesis test for a single proportion
   - Confidence Interval and Hypothesis Test for Two Proportions
   - Hypothesis test for multiple proportions
   - Statistics for contingency tables
3. Summary

Motivation

- So far we have assumed the normal distribution.
- But many relevant statistics have complicated distributions. For example:
  - The median
  - Quantiles in general
  - Any non-linear function of one or more (random) variables
- For the mean, we have learned that CLT (Central Limit Theorem) applies to *large* samples (but what if the sample is small and *not* normally distributed?).
- We lack tools when the assumptions for our tests are not met.
- **One solution**: Simulation and bootstrapping.

Bootstrapping

Bootstrap = pulling oneself up by the bootstraps

There are two versions of bootstrapping:

1. Parametric bootstrap: simulate repeated samples from the assumed (and estimated) distribution.
2. Non-parametric bootstrap: simulate repeated samples directly from the data.

Confidence interval for any sample statistic (incl. μ)

Method 4.7: Confidence interval for any θ by parametric bootstrap

Assume we have actual observations $y_1, \ldots, y_n$, and that they come from some probability distribution f (pdf).

1. Simulate $k \times n$ observations from the assumed pdf (with $\mu = \bar{x}$). [a]
2. Calculate the estimate $\hat{\theta}$ for each of the k samples, $\hat{\theta}_1^*, \ldots, \hat{\theta}_k^*$.
3. Find the $\alpha/2$- and $(1-\alpha/2)$-quantiles in $\hat{\theta}_1^*, \ldots, \hat{\theta}_k^*$, so that we get a $(1-\alpha)$-confidence interval: $\left[q^*_{\alpha/2}, q^*_{1-\alpha/2}\right]$

[a] Other parameters in the distribution should also match the data as well as possible

And the footnote...

"Other parameters in the distribution should also match the data as well as possible"

- For the normal distribution, choose μ and σ to match the sample's $\bar{x}$ and s.
- Some distributions have more than one parameter
- Generally, one should use the so-called *maximum likelihood* approach to match the distribution to the sample data.

Confidence interval for any sample statistic (comparison) $\theta_1 - \theta_2$ (incl. $\mu_1 - \mu_2$) from two samples

Assume we have actual observations $y_{1,1}, \ldots, y_{1,n_1}$, and $y_{2,1}, \ldots, y_{2,n_2}$, that these come from probability distributions f_1 and f_2. (The distributions are assumed independent)

1. Simulate k groups of 2 samples with n_1 and n_2 observations, respectively, from the assumed distributions, with means set to $\hat{\mu}_1 = \bar{y}_1$ and $\hat{\mu}_2 = \bar{y}_2$.
2. Calculate the difference between the sample statistics in each of the k samples: $\hat{\theta}^*_{y_1 1} - \hat{\theta}^*_{y_2 1}, \ldots, \hat{\theta}^*_{y_1 k} - \hat{\theta}^*_{y_2 k}$.
3. Find the $\alpha/2$- and $(1-\alpha/2)$-quantiles in these, $q^*_{\alpha/2}$ and $q^*_{1-\alpha/2}$, to obtain a $(1-\alpha)$-confidence interval: $\left[q^*_{\alpha/2},\, q^*_{1-\alpha/2}\right]$

Non-parametric bootstrapping: An overview

We do *not* assume any distribution!

Two methods for confidence intervals are provided:

	With one sample	With two samples
Any sample statistic	Method 4.15	Method 4.17

Confidence interval for any sample statistic θ (incl. μ) from one sample

We do *not* assume any distribution! This imply that we use the data itself.

Method 4.15: Confidence interval for any sample statistic θ by non-parametric bootstrapping

Assume we have observed $y_1, \ldots, y_n$.

1. Simulate k samples of size n by random sampling (with replacement) from the observed data (*re-sampling*).
2. Calculate the estimate $\hat{\theta}$ for each of the k samples: $\hat{\theta}_1^*, \ldots, \hat{\theta}_k^*$.
3. Find the $\alpha/2$- and $(1-\alpha/2)$-quantiles of these to obtain a $(1-\alpha)$ confidence interval: $\left[q^*_{\alpha/2},\, q^*_{1-\alpha/2}\right]$

Confidence interval for $\theta_1 - \theta_2$ (including $\mu_1 - \mu_2$) by non-parametric bootstrapping from two samples

Method 4.17: Confidence interval for $\theta_1 - \theta_2$ by non-parametric bootstrapping from two samples

Assume we have observations $y_{1,1}, \ldots, y_{1,n_1}$ and $y_{2,1}, \ldots, y_{2,n_2}$.

1. Draw k pairs of bootstrap samples with n_1 and n_2 observations from the respective samples (by random sampling with replacement).
2. Calculate the difference between the estimates in each of the k pairs of bootstrap samples:
   $\hat{\theta}^*_{y_1 1} - \hat{\theta}^*_{y_2 1}, \ldots, \hat{\theta}^*_{y_1 k} - \hat{\theta}^*_{y_2 k}$.
3. Find the $\alpha/2$- and $(1-\alpha/2)$-quantiles of these, $q^*_{\alpha/2}$ and $q^*_{1-\alpha/2}$, to obtain a $(1-\alpha)$ confidence interval: $\left[q^*_{\alpha/2},\, q^*_{1-\alpha/2}\right]$

Bootstrapping: An overview

We have seen 4 not so different method boxes

1. With or without distribution assumptions (parametric or non-parametric)
2. Analyses with one or two samples (one or two groups)

Note:

Means are also included in *random sample functions*. That is, these methods can also be applied for analyses beyond means!

Overview

1. Bootstrap
   - Parametric bootstrap
   - Non-parametric bootstrapping
2. Inference for proportions
   - Random variable for proportion
   - Hypothesis test for a single proportion
   - Confidence Interval and Hypothesis Test for Two Proportions
   - Hypothesis test for multiple proportions
   - Statistics for contingency tables
3. Summary

Different analyses and data types

Means in quantitative data

- Hypothesis test for a single mean based on one sample
- Hypothesis test for two means based on two samples
- Hypothesis test for multiple means based on several samples (coming later).

Today: Proportions in qualitative data

- Hypothesis test for a single proportion based on one sample.
- Hypothesis test for two proportions based on two samples.
- Hypothesis test for multiple proportions based on several samples.

Estimation of proportions

- We define the random variable P as the number of "successes" (Y) out of a total (n):

$$P = \frac{Y}{n}$$

- From sample data with y "successes" (sample size n), we estimate the proportion as:

$$\hat{p} = \frac{y}{n}$$

Note:

- $P \in [0; 1]$.
- p is the "true" population probability of a "success".

Binomial distribution

The number of "successes" (Y) follows a binomial distribution with the density function:

$$f(y;n,p) = \binom{n}{y} p^y (1-p)^{n-y}$$

Mean and variance in the binomial distribution

$$\begin{aligned} \mathbf{E}[Y] &= np \\ \mathbf{V}[Y] &= np(1-p) \end{aligned}$$

Mean and variance for proportions

Mean and variance for the proportion P:

$$
\begin{aligned}
\mathbf{E}[P] = \mathbf{E}\left[\frac{Y}{n}\right] &= \frac{np}{n} = p \\
\mathbf{V}[P] = \mathbf{V}\left[\frac{Y}{n}\right] &= \frac{1}{n^2}\mathbf{V}[Y] = \frac{p(1-p)}{n}
\end{aligned}
$$

Thus, we can define:

$$\sigma_P = \sqrt{\frac{p(1-p)}{n}}$$

Note:

σ_P is largest when $p = 1/2$.

For large n we approximately have

$$P \sim N(p, \sigma_P^2)$$

Confidence interval for a single proportion

Method 7.3

If the sample is large, then the $(1-\alpha)$-confidence interval for p is given by:

$$\hat{p} \pm z_{1-\alpha/2}\, \sigma_P$$

In practice, $\hat{p}$ is substituted for p in the formula $\sigma_P = \sqrt{p(1-p)/n}$

How?

This follows from approximating the binomial distribution with the normal distribution.

Rule of thumb

Assume $X \sim \text{bin}(n,p)$. The normal distribution is a good approximation for the binomial distribution if np and $n(1-p)$ (expected number of successes and failures) are both at least 15.

Margin of error (ME)

Margin of error

at a $(1-\alpha)$-confidence level is:

$$ME = z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

where we estimate p with $\hat{p} = \frac{x}{n}$.

Margin of error:

- Corresponds to half the width of the $(1-\alpha)$-confidence interval.
- Describes the expected precision (minimum desired precision) of the estimate $\hat{p}$.

Precision and sample size

Experiment planning:

How large does the sample size need to be to achieve a given precision?

Method 7.13

If you want an expected (given) margin of error (ME) in a $(1-\alpha)$-confidence interval, the required sample size is:

$$n = p(1-p)\left(\frac{z_{1-\alpha/2}}{\text{ME}}\right)^2,$$

where p (worst case $p = 1/2$) is a reasonable guess.

Example in Python

- Go to today's Python notebook in VS Code
  - "Example: Normal approximation of binomial distribution"

Steps in a hypothesis test – Overview

1. Formulate the null hypothesis and choose a significance level α.
2. Calculate the observed test statistic.
3. Calculate the p-value from the observed test statistic and the relevant distribution.
4. Compare the p-value with the significance level α and conclude.

Alternatively: Compare the observed test statistic with critical values and conclude.

Hypothesis test for a single proportion

We consider a null and alternative hypothesis for a single proportion p and choose a significance level α:

$$
\begin{aligned}
H_0 &: \quad p = p_0, \\
H_1 &: \quad p \neq p_0.
\end{aligned}
$$

As usual, reject H_0 or accept H_0.

Hypothesis test: Test statistic

Theorem 7.10 and Method 7.11

If the sample is large enough ($np_0 > 15$ and $n(1-p_0) > 15$), we use the test statistic:

$$z_{\text{obs}} = \frac{y - np_0}{\sqrt{np_0(1-p_0)}} = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

Under the null hypothesis, the test statistic approximately follows a standard normal distribution.

Find the p-value (evidence against the null hypothesis):

- $2P(Z > |z_{\text{obs}}|)$

Example in Python

- Go to today's Python notebook in VS Code
  - "Example: probability of rolling 6"

Confidence Interval for the Difference of Two Proportions

Method 7.15

$$(\hat{p}_1 - \hat{p}_2) \pm z_{1-\alpha/2} \cdot \hat{\sigma}_{\hat{p}_1 - \hat{p}_2}$$

where

$$\hat{\sigma}_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Rule of Thumb

Both $n_i\hat{p}_i \geq 10$ and $n_i(1-\hat{p}_i) \geq 10$ for $i = 1, 2$.

Hypothesis Test for the Difference of Two Proportions - Method 7.18

Hypothesis Test for Two Proportions

When comparing two proportions (shown here for a two-sided alternative hypothesis):

$$\begin{aligned} H_0: &\quad p_1 = p_2, \\ H_1: &\quad p_1 \neq p_2. \end{aligned}$$

Use the test statistic

$$z_{\text{obs}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}, \quad \text{where} \quad \hat{p} = \frac{y_1 + y_2}{n_1 + n_2}$$

Example 2

Is there a relationship between birth control pill use and the risk of heart clots?

A study (USA, 1975) investigated the association between birth control pill use and the risk of heart clots.

	Heart Clot	No Heart Clot
Pill Users	23	34
Non-Pill Users	35	132

Investigate whether there is an association between birth control pill use and the risk of heart clots. Use a significance level of $\alpha = 5\%$.

Example 2 – Continued

In a study (USA, 1975), the association between birth control pill use and the risk of heart clots was investigated.

	Heart Clot	No Heart Clot
Pill Users	23	34
Non-Pill Users	35	132

Estimates in each sample

$$\hat{p}_1 = \frac{23}{57} = 0.4035, \;\; \hat{p}_2 = \frac{35}{167} = 0.2096$$

Pooled Estimate:

$$\hat{p} = \frac{23+35}{57+167} = \frac{58}{224} = 0.2589$$

Example 2

- Go to today's Python notebook in VS Code
  - "Example: Contraceptive pills and risk of blood clots"

Hypothesis test for multiple proportions

Comparison of c proportions

In some cases, you may be interested in assessing whether two or more binomial distributions have the same parameter p, i.e., testing the null hypothesis:

$$H_0 : \quad p_1 = p_2 = ... = p_c = p$$

against the alternative hypothesis that these proportions are not equal (i.e., at least one is different).

Hypothesis test for multiple proportions

Table of observed counts for c samples:

	Sample 1	Sample 2	...	Sample c	Total
Success	y_1	y_2	...	y_c	y
Failure	$n_1 - y_1$	$n_2 - y_2$	...	$n_c - y_c$	$n - y$
Total	n_1	n_2	...	n_c	n

Common (average) estimate:

Under the null hypothesis, the estimate for p is:

$$\hat{p} = \frac{y}{n}$$

Hypothesis test for multiple proportions

Common (average) estimate:

Under the null hypothesis, the estimate for p is:

$$\hat{p} = \frac{y}{n}$$

"Use" this common estimate in each group:

If the null hypothesis is true, we expect the jth group to have e_{1j} successes and e_{2j} failures, where

$$e_{1j} = n_j \cdot \hat{p} = \frac{n_j \cdot y}{n}$$

$$e_{2j} = n_j(1 - \hat{p}) = \frac{n_j \cdot (n - y)}{n}$$

Hypothesis test for multiple proportions

Table with the *expected* counts in the c samples:

e_{ij}	Sample 1	Sample 2	...	Sample c	Total
Success	e_{11}	e_{12}	...	e_{1c}	y
Failure	e_{21}	e_{22}	...	e_{2c}	$n-y$
Total	n_1	n_2	...	n_c	n

General formula for calculating expected values in contingency tables:

$$e_{ij} = \frac{(\text{Row total } i) \cdot (\text{Column total } j)}{\text{total}}$$

Calculation of the test statistic - Method 7.20

The test statistic is

$$\chi^2_{\text{obs}} = \sum_{i=1}^{2} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

where o_{ij} is the *observed* count in cell (i, j) and e_{ij} is the *expected* count in cell (i, j).

Find p-value or use critical value – Method 7.20

Sampling distribution of the test statistic (under H_0):

χ^2 distribution with $(c-1)$ degrees of freedom (approximate)

Method with critical values:

If $\chi^2_{\text{obs}} > \chi^2_{1-\alpha}(c-1)$, then reject the null hypothesis.

Rule of thumb for validity of the test:

All expected values $e_{ij} \geq 5$.

Example 2 – continued

The *observed* values o_{ij}

Observed	Blood clot	No blood clot
Pills	23	34
No pills	35	132

Example 2 – continued

Use the "rule" for expected values four times, i.e.:

$$e_{22} = \frac{167 \cdot 166}{224} = 123.76$$

The *expected* values e_{ij}:

Expected	Blood clot	No blood clot	Total
Pills	14.76	42.24	57
No pills	43.24	123.76	167
Total	58	166	224

Example 2 – continued

Test statistic (include all cells):

$$\chi^2_{\text{obs}} = \frac{(23-14.76)^2}{14.76} + \frac{(34-42.24)^2}{42.24} + \frac{(35-43.24)^2}{43.24} + \frac{(132-123.76)^2}{123.76}$$
$$= 8.33$$

The critical value:

$\chi^2_{1-\alpha}(c-1)$ for $\alpha = 0.05$ and $c = 2$ (2 samples): 3.841

Conclusion:

Since $\chi^2_{\text{obs}} = 8.33 > 3.841$, reject the null hypothesis.

Example 2

- Go to today's Python notebook in VS Code
  - "Example: Contraceptive pills with χ^2"

Example 3: Analysis of a contingency table

A 3×3 table: 3 samples with 3 categorical outcomes

	4 weeks	2 weeks	1 week
Candidate I	79	91	93
Candidate II	84	66	60
Undecided	37	43	47
	$n_1 = 200$	$n_2 = 200$	$n_3 = 200$

Is the voting distribution the same?

$$H_0: \quad p_{i1} = p_{i2} = p_{i3}, \quad i = 1, 2, 3.$$

Another type of contingency table

A 3×3 table: 1 sample with two variables with 3 categorical outcomes:

	bad	average	good
bad	23	60	29
average	28	79	60
good	9	49	63

Is there independence between the classification criteria?

$$H_0: \; p_{ij} = p_{i\cdot} p_{\cdot j}$$

Test statistic – regardless of table type: Method 7.22

In a contingency table with r rows and c columns, the test statistic is:

$$\chi^2_{\text{obs}} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

where o_{ij} is the observed count in cell (i, j), and e_{ij} is the *expected count* in cell (i, j) (under the null hypothesis).

General formula for calculating expected values in contingency tables:

$$e_{ij} = \frac{(\text{Row total } i) \cdot (\text{Column total } j)}{\text{total}}$$

Find p-value or use the critical value - Method 7.22

Sampling distribution for the test statistic under H_0:

χ^2-distribution with $(r-1)(c-1)$ degrees of freedom.

Method with the critical value:

If $\chi^2_{\text{obs}} > \chi^2_{1-\alpha}$ with $(r-1)(c-1)$ degrees of freedom, then reject the null hypothesis.

Rule of thumb for validity of the test:

All expected values $e_{ij} \geq 5$.

Example 3

Does the distribution change "significantly" over time?

- Go to today's Python notebook in VS Code
    - "Example: Candidate votes over time"

Overview

Today: Proportions (Proportions)

When the outcome/interest variable y_i is **binary** (yes/no, success/failure, 0/1)

- Proportion in a group: $\hat{p}$
- Relevant null hypothesis is often $p_0 = 0.50$ (not zero!)
- Comparison of proportions in two or more groups
- (Not included in this course: Proportion as a function of explanatory variable, logistic regression)

When the outcome/interest variable y_i is a category with > 2 groups

- Discrete distribution between the groups (one proportion in each group)
- Comparison of distribution, e.g., over time or for different "exposures".

Overview

1. Bootstrap
   - Parametric bootstrap
   - Non-parametric bootstrapping
2. Inference for proportions
   - Random variable for proportion
   - Hypothesis test for a single proportion
   - Confidence Interval and Hypothesis Test for Two Proportions
   - Hypothesis test for multiple proportions
   - Statistics for contingency tables
3. Summary