Statistics

Camilo A. García Trillos

Market risk and portfolio theory

Hypothesis testing

Example: Claim (H_0): The random variable Z is normal with mean 1 and variance 4.

We observe $N = 10^6$ (one million) i.i.d samples of Z, with sample average 1.05.

Question: Is my initial claim reasonable given the observed data?

Hypothesis testing

Example: Claim (H_0): The random variable Z is normal with mean 1 and variance 4.

We observe $N = 10^6$ (one million) i.i.d samples of Z, with sample average 1.05.

Question: Is my initial claim reasonable given the observed data?

Refined question: Under the assumption, claim is **unreasonable** if it is *very unlikely* to observe an even greater distance from the mean.

Hypothesis testing

Solution: By the central limit theorem, we have

$$\hat{Z}_1 := \sqrt{N}\frac{(\hat{Z} - \mathbb{E}[Z])}{\sqrt{\mathrm{var}[Z]}} \xrightarrow{d} \mathcal{N}(0, 1)$$

where $\hat{Z}$ is the sample mean r.v. In our case,

$$P[|\hat{Z}| > \hat{z}] = \mathbb{P}[|\hat{Z}_1| > a]$$

where

$$a = \frac{\sqrt{10^6} \times (1.05 - 1)}{\sqrt{4}} = 25.$$

Hence, $P[|\hat{Z}| > \hat{z}] \approx 6.113 \times 10^{-138}$!!!

Thus, either our claim is false or we are observing an extremely unlikely event. We *reject the null hypothesis*.

Hypothesis testing

Let's revisit the question, this time assuming we observe $N = 100$ i.i.d. samples of Z with sample average 1.05 as before. We get

$$a = \frac{\sqrt{10^2} \times (1.05 - 1)}{\sqrt{4}} = 0.25$$

and thus, $P[|\hat{Z}| > \hat{z}] \approx 0.8026$.

This does not seem as an unlikely event any more.

Question: Can we conclude that the null hypothesis holds?

Hypothesis testing

		Decision	
		Retain null	Reject null
Truth in population	True	Correct: $(1 - \alpha)$	Type I error: α
	False	Type II error	Correct

- Unless an alternative is considered, we focus on obtaining evidence to reject the null assumption (small type I error), but not on obtaining evidence to support it

Hypothesis testing

		Decision	
		Retain null	Reject null
Truth in population	True	Correct: $(1-\alpha)$	Type I error: α
	False	Type II error: β	Correct (power) : $(1-\beta)$

- Unless an alternative is considered, we focus on obtaining evidence to reject the null assumption (small type I error), but not on obtaining evidence to support it
- If an alternative assumption is available, we can also control the type II error by choosing the number of samples and statistics.

Reminder of hypothesis testing

To summarise: To perform the test

1. State the hypotheses (null hypothesis, H_0)
2. Set the criteria for decision:
   - Estimator
   - Reference probability for rejection α
   - If alternative assumption available fix also β
   - Type of test (two-tailed, left-tailed or right-tailed)
3. Compute the test statistic and its p-value
4. Make a decision: if p-value is smaller than reference, reject the null hypothesis.

Some examples of tests

- **Wald test**:
  Asymptotic Gaussian statistic $\hat{Z}$ as above. Used to compare scalars.
  In this case, a hypothesis test of level α is equivalent to checking if the null value is in a $1 - \alpha$ confidence interval.
  Ex: means, medians, probabilities in binomial distributions.
- **Likelihood ratio test**: Useful for $H_0 : \theta \in \Theta_0$.
  In the case where Θ_0 is of the form 'the last ℓ entries are fixed', the statistic

$$2 \log \left(\frac{\sup_{\theta \in \Theta} \mathcal{L}(\theta)}{\sup_{\theta \in \Theta_0} \mathcal{L}(\theta)} \right)$$

  is asymptotically χ^2_ℓ.

Some examples of tests (cont.)

- χ**-test for multinomial**:
  Useful to test for multinomial distribution. The statistic

$$\sum_{j=1}^{k} \frac{(X_k - np_{0j})^2}{np_{0j}}$$

  is asymptotically χ^2_{k-1}.
  This test can also be adapted for goodness-of-fit and independence.

Linear regression (multidimensional case)

Consider the problem of selecting coefficients α, β so that

$$Y = \boldsymbol{F} \cdot \boldsymbol{\beta} + \alpha + \epsilon = \alpha + \beta_1 F_1 + \ldots + \beta_K F_K + \epsilon$$

where $\mathbb{E}(|\epsilon|^2)$ is minimal.

It can also be written

$$Y = \bar{\boldsymbol{F}} \cdot \bar{\boldsymbol{\beta}}$$

where

$$\bar{F} = \begin{pmatrix} 1 \\ \boldsymbol{F} \end{pmatrix}; \qquad \bar{\beta} = \begin{pmatrix} \alpha \\ \boldsymbol{\beta} \end{pmatrix}$$

Linear regression (multidimensional case)

Treating F as a matrix (of samples), we solve the problem (without intersect) by choosing

$$\hat{\beta} = (F^\top F)^{-1} F^\top Y$$

provided that this is well-defined. Note that

$$\mathbb{V}(\hat{\beta}|F) = \sigma^2 (F^\top F)^{-1}$$

Model selection

- AIC (Akaike Information Criterion): Minimise $|S| - \ell_S$ where ℓ_S is log-likelood at the MLE.
- BIC (Bayesian Information Criterion): Minimise $\frac{|S|}{2}\log(n) - \ell_S$

Generalization

GLS: If the matrix $F^\top F$ is ill-conditioned, the estimators will be poor. Statistically the estimators will not be 'efficient'.
Statistically, this occurs when there is heteroscedasticity and errors are not independent. The following estimator would be ideal:

$$\hat{\beta} = (F^\top \Sigma^{-1} F)^{-1} F \top \Sigma^{-1} Y$$

where $\Sigma = \mathrm{cov}(\epsilon | F)$. Approximations are enough to improve the OLS estimation.