TP : Confidence intervals

1 Central Limit Theorem

The Central Limit Theorem (CLT) is a fundamental concept in statistics that asserts the following : When you have a sufficiently large number of independent and identically distributed (i.i.d.) random variables, the distribution of their sample mean will closely approximate a normal distribution. This holds true regardless of the original distribution of the individual random variables, as long as that original distribution has a finite mean and variance.

To illustrate the Central Limit Theorem by simulating it with a uniform distribution, you can utilize Python and follow these steps :

- Generate Random Samples : Create m random samples from a uniform distribution. Python's `numpy` library is a useful tool for this purpose.
- Calculate Sample Means : Compute the means of these m random samples, denoted as $\bar{x}_i$ for $1 \leqslant i \leqslant n$. According to the CLT, the distribution of these sample means, $\bar{x}_i$, should closely resemble a normal distribution.
- Visualize the Result : To gain a better understanding, craft a histogram of the sample means. This visualization will help depict the approximate normality.
- Repeat for Different m : Perform this simulation for various values of m to observe how the approximation to a normal distribution improves as the sample size increases.

2 Confidence Intervals

Use `norm.ppf` and `t.ppf` in the following exercises :

1) Let `S=array([16.621, 17.416, 19.147, 9.745, 13.082, 16.125, 15.078, 4.676, 16.923, 21.853, 12.971, 10.198, 24.208, 18.028, 11.363, 10.702, 22.305, 13.151, 20.258, 13.949])`.
2) Give a point estimate for the mean.
3) Compute a confidence interval of level 99% for the mean of S assuming S was generated i.i.d. from a Gaussian distribution with an unknown mean and standard deviation.
4) Compute a confidence interval of level 95% for the mean of S assuming S was generated i.i.d. from a Gaussian distribution with a known mean and standard deviation.
5) (bonus) Recall the distribution of the variance and give an appropriate confidence interval for the variance assuming S was generated i.i.d. from a Gaussian distribution with an unknown mean and standard deviation.

3 Linear Models

In this section, we will focus on the `diabetes` dataset available in `sklearn`, specifically considering only the variable `bmi` and the response variable y. For the purposes of this analysis, we will treat this dataset as a problem with a single independent variable.

3.1 Confidence intervals for the coefficients

Start by visualizing the dataset, specifically examining the relationship between the `bmi` variable and the response variable y. Create a scatter plot to explore the data in these two dimensions. Then, follow these steps.

- **Estimation of Coefficients** : We will estimate the coefficients for the linear regression model using `sklearn`. After estimating the coefficients, we will plot the least squares regression line that represents the best-fitting linear relationship between `bmi` and y.

- **Confidence Intervals for the slope** : Additionally, we will compute confidence intervals for these coefficient estimate $\hat{\theta}_0$ and $\hat{\theta}_1$ using the expressions seen in class. Compare the solution with `statsmodel` is you want.

3.2 Bootstrap Resampling

We will employ a bootstrap resampling technique to assess the stability and variability of our linear regression model. Here are the steps :

- **Bootstrap Samples** : Generate 100 bootstrap samples from the original dataset. Each bootstrap sample will be randomly selected with replacement from the observed data.
- **Regression Lines from Bootstrap Samples** : Fit linear regression models to each of the 100 bootstrap samples. These models will provide a range of potential regression lines that capture the variability in the data.
- **Plotting Results** : Plot the regression lines obtained from the bootstrap samples on the same graph, ensuring transparency to distinguish different lines. This visualization will give insight into the range of possible relationships between `bmi` and y.
- **Confidence Intervals from Bootstrap** : Calculate and provide the confidence interval for the regression coefficients based on the results obtained from the bootstrap samples. This interval quantifies the uncertainty associated with our coefficient estimates.

3.3 Prediction Intervals

In this section, our primary focus is on calculating the prediction interval for a mean response in the context of a linear regression problem. Specifically, we will consider the same regression problem introduced in the previous section, which utilizes only the `bmi` feature.

The objective at hand is to incorporate the prediction interval into the existing plot. In the following paragraphs, we will delve into what a prediction interval represents and outline the steps to compute it.

A prediction interval within the realm of linear models signifies a statistical range that provides an estimate of where a future observation is likely to fall with a specified level of confidence. To calculate a prediction interval for a simple linear regression model with just one independent variable, please follow these steps :

1) **Fit a Simple Linear Regression Model** : First, fit a simple linear regression model to your data. The model has the form :

$$\hat{y} = \theta_0 + \theta_1 x$$

Where :

- $\hat{y}$ is the predicted value of the dependent variable (the response).
- θ_0 is the estimated intercept.
- θ_1 is the estimated coefficient for the independent variable (predictor), denoted as x.

2) **Calculate the Point Prediction** : For the observation you want to predict (let's call it x_0), plug its value into the regression equation to calculate the point prediction :

$$\hat{y}_0 = \theta_0 + \theta_1 x_0$$

3) **Calculate the Standard Error of Prediction (SE)** : The standard error of prediction accounts for the uncertainty in your prediction and is calculated as follows :

$$SE = \sqrt{MSE * \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)}$$

Where :

- MSE is the mean squared error of the regression model.
- n is the number of observations.
- x_0 is the value of the independent variable for the observation.

- $\bar{x}$ is the mean of the independent variable in the training data.
- x_i are the values of the independent variable in the training data.

4) **Choose the Confidence Level** : Decide on the desired level of confidence for your prediction interval (e.g., 95% confidence, 99% confidence).
5) **Find the t-Statistic** : Find the critical value (t-statistic) from the t-distribution corresponding to your chosen confidence level and degrees of freedom. The degrees of freedom are typically equal to the sample size minus 2 (for a simple linear regression with one predictor and one intercept).
6) **Calculate the Prediction Interval** : Calculate the lower and upper bounds of the prediction interval for all x_0 using the following formula :

$$\text{Lower Bound} = \hat{y}_0 - t_{\alpha/2} \times SE$$

$$\text{Upper Bound} = \hat{y}_0 + t_{\alpha/2} \times SE$$

Where :

- $t_{\alpha/2}$ is the critical value from the t-distribution corresponding to the chosen confidence level and degrees of freedom.
- $\alpha/2$ is the significance level divided by 2.

7) **Plot the Prediction Interval** : Finally, plot the lower and upper bounds of the prediction interval for all x_0 in the range of X.

These steps will give you a prediction interval for the mean response for the simple linear regression case with one independent variable, where θ_0 represents the intercept, θ_1 represents the coefficient for the independent variable, and x_0 is the value of the independent variable for the observation.

3.4 Feature Selection

In linear models, when the coefficient $\theta_j = 0$ for $j > 0$, it indicates that the feature (column) X_j has no impact on predicting the regression problem for y. In this section, we will create a program to independently test whether each feature has a significant effect on the regression. This sequence of tests can serve as a foundational step in feature subset selection methods. The process unfolds as follows :

1) We define a vector of covariates with intercept $\tilde{X} = (\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_{10})$.
2) for each variable $\tilde{X}_k$, $k = 1, \ldots, 11$, we consider the linear model that consists on just the feature X_j and the observation y

$$\mathbf{y} \simeq \theta_k \mathbf{x}_k$$

3) we test whether its regression coefficient equals zero, *i.e.,*

$$H_0 : \theta_k = 0$$

using the statistic $\frac{\hat{\theta}_k}{SE(\hat{\theta}_k)}$. Display the statistics for each feature.
4) Compute the p-values for each feature, and keep the one possessing the smallest p-value.
5) For which features will we reject the null hypothesis ?
6) Think on how to use this to propose a greedy algorithm for Feature Subset Selection.

Note : Remember how we derive the test of null effect for coefficient θ_j.
Let $s_x^2 = n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ and $\frac{(\hat{\theta}_1 - \theta_1^*)\sqrt{ns_x^2}}{\hat{\sigma}*} \sim T_{n-2}$ (or equivalently $\frac{(\hat{\theta}_1 - \theta_1^*)}{SE(\hat{\theta}_1)} \sim T_{n-2}$)

- Under H_0, which assumes that $\theta_1^* = 0$, we have $\frac{(\hat{\theta}_1)}{SE(\hat{\theta}_1)} \sim T_{n-2}$
- Let $t = \frac{(\hat{\theta}_1)}{SE(\hat{\theta}_1)}$ be our statistic
- $t_{\alpha/2}$ be the quantile of degree $\alpha/2$ for a T-student distribution.
- If $|t| > t_{\alpha/2}$, reject H_0, i.e., we have not strong argument to conclude that $\theta_k = 0$.