* Dealing with missing data:

Option 1: delete the whole raw of data.

→ **drawback:**

(i) Assume we have a dataframe with 10,000 raws and 10 columns. And in all the raws 5 columns have possible null values. If these 6 columns have 2,000 missing values and we remove them, then almost 20% of data is reduce.

what happend if we have 20 columns? The number of null value also increase, so we may reduce our data set 40% or more.!!

Option-2: Purify the Null values.

* Now option-2 definately more effective, because we have more varibility in the dataset, as the raw number increase.

✳ The Aim of Imputing Missing values:

- The aim of imputation is not exactly predit it.
- The aim is to complete the data so the entire data can be used, and the variance reduce and accuracy increase.

✳ Type of missing data:

(1) Missing completely at random (MCAR):

Missing comptely at random is a mechanism in missing data analysis where missingness occurs randomly and independently of any observed or unobsereved variables.

In statistical term: the probability that a missing value occure is constant. The occurrence is independent from any other varibles.

when data is missing completely at random, the analysis of the observed data can be proceed without bais if missing data is appropiately handled.

2. Missing at random (MAR):

In MAR, whether a data point is missing or not depends on the value of other observed variables in the dataset, but not on the value of the variable that is missing.

In statistical terms: Missing at random is a mechanism in data analysis where the probability of missing data depends on the observed data, but not on the unobserved data. The probability is not constant, but can be fully explained using variable of the dataset.

3. Missing not at random (MNAR):

In MNAR, wheather a data point is missing or not, is depends on the value of unobserved variables, which are missing themselves. That means, this mechanism introduces a systematic pattern in the missing data that can't be explained by the observed data.

In statistical term, MNAR is the mechanism in missing data analysis where the probability of the missing data depends on the unobserved data, even after taking into account of observed data.

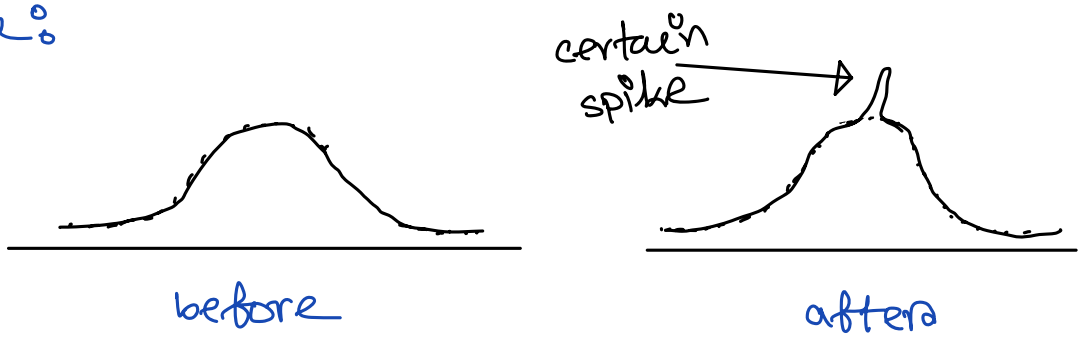
✳ For MCAR and MAR, omitting data should not Bias the results, but we will loss prediction power.

✳ Univariate Imputation Methods (Fixed Methods):

- Mean replacement for each variable independently, calculate the mean of the non-missing values and set the missing values equal to that mean.
- For example, pulse rate has 1437 missing values, and the mean pulse rate of the known values is 73.56.
- Mean replacement will replace the missing values with 73.56.
- This method is not good because the standard deviation of the imputed variable will decrease. As a result, the correlation with other variable will also decrease.

- For example, the standard deviation is 12.15 but after mean replacement the standard deviation drop to 11.25.

- Graphical representation of the distribution of the data:



- Missing as a category: For categorical variable add a new level called missing. For example of gender define three variable: male, female, and missing.

- **Using logic rule:** Some time a variable can be imputed exactly with high accuracy.

  → **Example-1**, if a person's height and BMI is known, its easy to calculate the weight using BMI-index.

  → **Example-2**, if a variable is number of pregnancies and a positive number is given then a missing value for gender can be imputed as female.

→ Example—3, only 1% of breast cancer cases are in males. This means, if there is an indication of current or previous breast cancer, it is acceptable that the missing gender is female.

✳ Univariate Imputation Methods (Random Methods)

⊙ Imputations using Mean/Variance simulations :

→ Suppose the variables X_1 has missing values.

→ Compute the mean $\bar{X}_1$ and std. deviation S_{X_1} of the known values.

→ Now replace the missing values with the simulated data from $\sim N(\bar{X}_1, S^2_{X_1})$ distribution.

→ This method provide the missing values roughly normal distributed. If the variable is very non-normal then after imputation the variable become close to normal distribution.

→ The disadvantage with this and all univariate methods, is that it doesn't consider any co-dependencies with other variables.

⊙ Direct Random Sample:

→ Sample the non-missing values with replacement, to fill in the missing values.

→ **steps:**

1. Create two subset from the original data. where first subset contains non-missing values and second subset contains missing values.

   i.e. $S = \{3, NaN, 4, 9, NaN, 6, 8, NaN\}$

   $S_1 = \{3, 4, 9, 6, 8\}$ $S_2 = \{NaN, NaN, NaN\}$.

2. Then it randomly select n-th number of data from the non-missing list, where n-is the number of missing values.

3. Finally, it sets the newly available values into the original dataset.

→ This approach is good when the missingness is independent of all other variable.

→ Compare the mean/variance simulation method, it preserve the original distribution properties of the variable, and can be used for all data types.

→ It is often used as a first step in multivariate methods.

✳ Multivariate Imputation for one variable:

- With most data, there is co-dependency (correlation) between some variables, which can and should be incorporated into our imputation method.
- For example, if we know a person's height, age, and gender, then we can impute the weight much more accurately than just sampling the non-missing values of weight.
- We will assume from here the dataset consists entirely of continious variables. Allowing other types of variables is more complex but follows similar methods.

✳ Multivariate Regression:

- We can take the elements with data and fit a regression model to predict the missing values for another variable.
- The type of regression can be any regression method but is often multiple linear regression.
- For example, we assume we have complete data (n= 10,000) for age, race, gender, and BMI. The variable pulse has 1437 missing values.

→ Now, we can fit a linear model

Pulse ~ BMI + age + race + gender

Using 8563 elements with complete data.

→ The predicted missing values should then be simulated with an error term.

→ If $\hat{y}_i$ is the predicted value from the regression model then the imputed value should be:

$$\tilde{y}_i = \hat{y}_i + \varepsilon_i ,$$

where ε_i is a random normal simulation with model residual variance.

→ As mean/variance simulation, simulating the error term avoid a reduction in the variance of that variable.

→ The aim is not to get the best prediction for missing value but to preserve the overall properties of the complete dataset.

→ The regression method is usually used for continuous variables.

→ For binary variables logistic regression may used.

→ For factor variable with more than two levels use a tree model of other type of simple classifier.

✳ When more than one variable has missing values:

- The multivariate regression method easily adapts itself to imputing multiple variables, by looping over the variables with missing values.

- suppose, $X_1, X_2, \cdots X_k$ are variables with some missing values and $X_{k+1}, X_{k+2}, \cdots, X_p$ are complete.

  → Fit the regression $X_1 \sim X_{k+1} + X_{k+2} + \cdots + X_p$ and impute all missing values in X_1 including a simulated error term.

  → Then fit the regression $X_2 \sim X_1 + X_{k+1} + X_{k+2} + \cdots + X_p$ and impute all the missing values in X_2 with simulated error term.

  → Continue until all variables $X_1, \cdots, X_k$ have been completed.

- The problem of this method is that if X_1 is correlated with X_2 then this is not considered when imputing X_1, but is considered when imputing X_2.

- To solve this problem the following steps are used:

1. Firstly, used a simple univariate model like direct random sampling to complete each missing variable. Keep a records of the "positions" of the original missing values.

2. Fit the regression $X_1 \sim X_2 + X_3 + \cdots + X_p$, using all elements for which X_1 was not imputed.

3. Re-impute all the originally missing values in X_1 including a simulated error term.

4. Then fit the regression $X_2 \sim X_1 + X_3 + \cdots + X_p$, using all elements for which X_2 was not imputed, and re-impute.

5. Repeat for $X_3 \cdots X_k$.

6. Because the X_1 imputation were based on some direct random sampled values of $X_2, \cdots, X_k$, it is best to repeat this loop a couple more times.

✳ Imputing Missing values using GIBBS Sampling:

- From mean/variance sampling: Replace the missing values of X_1 with simulations from a normal $N(\hat{\mu}_1, \hat{\sigma}_1^2)$ distribution.

- The core disadvantage with this and all univariate methods, is that it does not consider the co-dependencies other variables.

- Gibbs sampling (GS) is a method which simulates the missing data by estimating $\hat{\mu}_1, \hat{\sigma}_1, \hat{\mu}_2, \hat{\sigma}_2, \cdots, \hat{\mu}_p, \hat{\sigma}_p$ simultaneously.

- For each GS iteration a slightly different value for each of these parameter estimates which reflects that estimate themselves have some variability.

✳ Gibbs sampling :

- There is a wide class of **simulation algorithm** called **Markov Chain Monte Carlo (MCMC)** simulation which used **Bayesian estimation.**

- These method allow us to **simulate parameter** values that comes from so called **Posterior distribution** of these parameters.

- Simulation from the **Posterior distribution** using MCMC is a **powerful tool** in Bayesian estimation.

- **Gibbs sampling** is the most common **MCMC algorithm.** Each time a **new parameter** value is **simulated** it **depends on the value of known data**, the **values** of **other parameters**, and **the other imputed values.**

- Examples :

```
> md.pattern(airquality)
    Wind Temp Month Day Solar.R Ozone
111    1    1     1   1       1     1  0
35     1    1     1   1       1     0  1
5      1    1     1   1       0     1  1
2      1    1     1   1       0     0  2
       0    0     0   0       7    37 44
```

(2) → Total missing in both solar.R and Ozone.

(7) → Total missing in column solar.R

(37) → Total missing in Ozone

Summary of md.pattern():

1. There are total 153 observation and 6 features.
2. Total missing elements are 44.
3. Among them 40 raws have only one missing data and 2 raws have 2 missing data.
4. These 44 missing values are distributed in Solar.R and Ozone columns. where there are 7 missing values are in Solar.R columns and rest 37 missing data are in Ozone columns.

Imputation Using Gibbs sampling:

```
# Imputation Using Gibbs sampling
tempData <- mice(airquality, m = 5, maxit = 50, meth = 'norm', seed = 500)
```

- dataset
- obtain '5' different imputation (relisation)
- Numbe of iteration
- method used to obtain the conditional distributions and updates, "norm" means "Bayesian linear Regression".

- To obtain complete dataset with 1st imputation:

```
# To obtain a full data set with the 1st imputation.
completedData <- complete(tempData,1)

md.pattern(completedData)
```

impute using 1st realisitions.

```
 /\     /\
{  `---'  }
{  O   O  }
==>  V <==  No need for mice. This data set is completely observed.
 \  \|/  /
  `-----'

    Ozone Solar.R Wind Temp Month Day
153     1       1    1    1     1   1 0
        0       0    0    0     0   0 0
```

- Suppose the original plan was to fit a multiple linear regression model to predict temperature using ozone, Solar.R, and wind.

→ Now we have complete data we can fit this regression model, but we have 5 realisitions. Rather than choose one of the 5, we can fit the model to each of the realisitions and "Pool" (average) the results. This should be done using pool function.

```
> # fitting model using imputed data
> modelFit1 <- with(tempData, lm(Temp ~ Ozone + Solar.R + Wind))
> summary(pool(modelFit1))
          term    estimate   std.error statistic       df      p.value
1 (Intercept) 71.27152600 2.959069951 24.085786 69.78416 1.544101e-35
2       Ozone  0.17429782 0.025314421  6.885317 46.61500 1.269689e-08
3     Solar.R  0.01038069 0.006968797  1.489595 91.15202 1.397828e-01
4        Wind -0.26838481 0.213902136 -1.254708 77.01801 2.133789e-01
```