Nama : Ghina Khoerunnisa
Program : Data Scientist Intern

Inferential Statistics

Inferensial Statistik adalah metode statistik yang mengambil data dari sampel untuk membuat kesimpulan pada populasi yang lebih besar. Ada banyak cara untuk memilih sampel dari populasi, tetapi secara umum, pengambilan sampel secara acak memungkinkan kita untuk memiliki keyakinan bahwa sampel tersebut mewakili populasi.

Probability Distribution:

- **Uniform ->** Setiap interval angka dengan lebar yang sama memiliki probabilitas yang sama. Untuk membuatnya dapat menggunakan uniform function dari scipy.stats module. uniform.rvs(size=1000, loc=10, scale=20) yang mana size merupakan jumlah variasi acak.
- **Normal ->** Distribusi normal memiliki kurva berbentuk lonceng (simetris) yang dijelaskan dengan mean μ dan standar deviasinya σ. Distribusi dengan mean 0 dan standard deviation 1 disebut standard normal distribution. Untuk membuatnya dapat menggunakan scipy.stats module yaitu norm.rvs(). Argumen loc sesuai dengan rata-rata distribusi, scale sesuai dengan standard deviation dan size sesuai dengan jumlah variasi acak.
- **Bernoulli ->** Hanya memiliki dua kemungkinan hasil, yaitu 1 (berhasil) dan 0 (gagal), dan dalam sekali percobaan. Untuk membuatnya dapat menggunakan scipy.stats module bernoulli.rvs(size=10000,p=0.6), yang mana p merupakan probabilitasnya.
- **Binomial ->** Hanya dua hasil yang mungkin, seperti sukses atau gagal di mana probabilitas keberhasilan dan kegagalan sama untuk semua percobaan. Parameter dari binom.rvs() adalah n dan p dimana n adalah jumlah total percobaan, dan p adalah probabilitas keberhasilan pada setiap percobaan.
- **Poisson ->** digunakan untuk memodelkan berapa kali suatu peristiwa terjadi dalam interval waktu. Misalnya jumlah pengunjung di restoran dalam suatu interval dapat dianggap sebagai proses Poisson. Untuk membuatnya dapat menggunakan scipy.stats module poisson.rvs() diman mu adalah meannya.
- **Exponential ->** Menggambarkan waktu antara peristiwa dalam Poisson point process. Misalnya peluang waktu yang dibutuhkan untuk 2 bus lewat (event nya sekali -> seperti bernoulli). Untuk membuatnya dapat menggunakan scipy.stats module expon.rvs().
- **Gamma ->** Gamma distribution adalah two-parameter family dari continuous probability distributions. Mirip seperti exponential tapi event nya lebih dari sekali. Untuk membuat gamma distributed random variable dapat menggunakan metode gamma.rvs() yang mempunyai parameter shape a sebagai argumennya. Ketika a bilangan bulat, gamma tereduksi menjadi distribusi Erlang, dan ketika a = 1 menjadi distribusi eksponensial.

Confidence Intervals, CI adalah jenis estimasi yang dihitung dari data statistik yang diamati. CI menghitung seberapa akurat mean sebuah sampel mewakili nilai mean populasi

sesungguhnya. Sehingga, CI adalah rentang antara dua nilai di mana nilai suatu sampel mean tepat berada di tengah-tengahnya. Rumus CI proportion atau mean adalah sebagai berikut:

$$Population\ Proportion\ or\ Mean\ \pm z-score * Stadard\ Error$$

Rumus standard error untuk proportion dan mean adalah

$$Standard\ Error\ For\ Population\ Proportion = \sqrt{(Population\ Proportion * \frac{(1 - Population\ Proportion)}{Number\ Of\ Observations})}$$

$$Standard\ Error\ For\ Mean = \frac{Standard\ Deviation}{\sqrt{Number Of\ Observations}}$$

Rumus CI yang sebelumnya dapat digunakan jika data kita berdistribusi normal. Jika data kita tidak terdistribusi secara normal maka dapat mengimplementasikan central limit theorem. Berdasarkan central limit theorem, pengambilan sampel dalam jumlah yang cukup dengan ukuran yang memadai akan menghasilkan distribusi sample means yang normal. Rumus CI menjadi sample mean ± (z-score * standard error) dimana standar errornya adalah standar deviasi populasi dibagi dengan jumlah sampel.

Hypothesis Testing, Hipotesis (anggapan dasar) adalah jawaban sementara terhadap masalah yang masih bersifat praduga karena masih harus dibuktikan kebenarannya. Hipotesis harus dapat diuji, baik dengan eksperimen atau observasi. Statement hipotesis yang baik dan benar memiliki kriteria yaitu merupakan sebab akibat (jika … maka …), terdapat independen dan dependent variable, dapat dites dengan eksperimen, survey, atau teknik scientific lainnya**,** berdasarkan research, dan memiliki kriteria desain. Hipotesis testing adalah cara menguji hasil survei atau eksperimen untuk melihat apakah hasilnya bermakna. Pada dasarnya menguji apakah hasilnya valid dengan mencari tahu kemungkinan bahwa hasil yang terjadi secara kebetulan. Tahap-tahapannya adalah:

- **Mencari null hypothesis**
- **Menetapkan null hypothesis**
- **Pilih jenis tes yang perlu dilakukan**
- **Support atau Reject null hypothesis**

Hasil uji hipotesis statistik harus diinterpretasikan agar kita dapat mulai membuat klaim. Statistical hypothesis test dapat mengembalikan nilai yang disebut p-value. Ini adalah kuantitas yang dapat digunakan untuk mengukur hasil tes: **reject or fail to reject**. Nilai umum yang digunakan untuk alpha adalah 5% atau 0,05.

- Jika p-value > alpha: Fail to reject null hypothesis (not significant result).
- Jika p-value <= alpha: Reject null hypothesis (significant result).

Signifikan dalam statistik adalah benar-benar berbeda atau nyata. Macam-macam Statistical Hypothesis Test adalah ztest, Normality Test (Shapiro-Wilk Test, D'Agostino's K^2 Test, Anderson-Darling Test), Correlation Test (Pearson's Correlation Coefficient, Spearman's Rank Correlation, Kendall's Rank Correlation Chi-Squared Test), Stationary Test (Augmented Dickey-Fuller Unit Root Test, Kwiatkowski-Phillips-Schmidt-Shin), Parametric Statistical Hypothesis Tests (Student's t-test, Paired Student's t-test, Analysis of Variance Test (ANOVA)), dan Nonparametric Statistical Hypothesis Tests (Mann-Whitney U Test, Wilcoxon Signed-Rank Test, Kruskal-Wallis H Test, Friedman Test).