# Inferential Statistics

Tujuan utama dari metode ini adalah untuk menarik kesimpulan dari sampel dan menggeneralisasikannya untuk populasi yang lebih besar. Oleh karena itu, kita perlu mencari sampel sedemikian rupa sehingga dapat mencerminkan populasi dengan benar.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.

Dalam banyak kasus, tidak mungkin mengukur atau memahami keseluruhan populasi;dalam hal ini, pengambilan sampel menyelamatkan kita.Akibatnya, kecil kemungkinannya nilai-nilainya sama ketika kita mencoba memperkirakan sifat-sifat populasi dari suatu sampel.

# Why you need inferential statistics

Ada 2 metode yang dapat kita gunakan untuk menghitung hasil:

* Collect data about each and every child.
* Use the data we have to calculate the overall average.

Metode pertama adalah tugas yang sangat sulit dan ribet.Jumlah usaha dan sumber daya yang dibutuhkan untuk menyelesaikan tugas ini akan sangat besar.

Metode kedua jauh lebih sederhana dan lebih mudah dilakukan.Tapi ada masalah.Kita tidak dapat menyamakan rata-rata yang kita peroleh dari limited data dengan seluruh populasi.Pertimbangkan kasus di mana anak-anak di daerah kita lebih tertarik pada olahraga sehingga jumlah jam yang mereka habiskan untuk menonton televisi jauh lebih sedikit daripada jumlah keseluruhan populasi.Bagaimana cara kita mencari rata-rata populasi?Di sinilah inferential statistics datang untuk membantu kita.

Inferential statistics membantu kita menjawab pertanyaan-pertanyaan berikut:

* Making inferences about a population from a sample
* Concluding whether a sample is significantly different from the population.Let’s look at the previous example where I pointed out that the sample is different from the population as the children are more interested in sports rather than watching television.
* If adding or removing a feature from a model will help in improving it.
* If one model is significantly different from the other.
* Hypothesis Testing.

# Probability Distribution

Probability distribution adalah fungsi di bawah teori dan statistik probabilitas - yang memberi kita seberapa besar kemungkinan hasil yang berbeda dalam sebuah eksperimen.Probability distribution menggambarkan peristiwa dalam hal probabilitasnya;dari semua kemungkinan hasil.Mari kita ambil distribusi probabilitas dari lemparan koin yang adil.Di sini, kepala mengambil nilai X = 0,5 dan ekor mendapat X = 0,5 juga.

Dua kelas dari distribusi semacam itu adalah discrete dan continuous.Yang pertama diwakili oleh probability mass function dan yang terakhir diwakili oleh probability density function.

# Random Variable

Variabel yang kemungkinan nilainya merupakan hasil numerik dari fenomena acak. Ada dua jenis variabel acak, diskrit dan kontinue.

## Discrete Random Variable

Variabel yang hanya mengambil sejumlah nilai berbeda yang dapat dihiting dan dengan demikian dapat dikuantifikasi. Misalnya, kita dapat menentukan variabel acak X sebagai angka yang muncul saat kita melempar dadu yang adil. X bisa mengambil nilai [1,2,3,4,5,6] sehingga variabel acak diskrit.

Probabiliy distribution dari discrete random variable adalah daftar probabilitas yang terkait dengan masing-masing kemungkinan nilainya. Disebut juga probability function atau probability mass function.

Untuk memiliki pengertian matematis, anggaplah variabel acak X dapat mengambil nilai k yang berbeda, dengan probabilitas bahwa X=xᵢ didefinisikan sebagai P(X=xᵢ)=pᵢ.Maka probabilitas pᵢ harus memenuhi hal-hal berikut:

* 0 < pᵢ < 1 for each ᵢ
* p₁+p₂+...+pₖ=1

Beberapa contoh dari discrete probability distributions adalah Bernoulli distribution, Binomial distribution, Poisson distribution, dll.

## Continuous Random Variable

Variabel yang mengambil jumlah kemungkinan nilai tak terbatas. Misalnya kita dapat menentukan variabel acak X sebagai tinggi siswa di kelas.

Dikenal sebagai probability density function adalah fungsi yang menggunakan nilai kontinu.

Probabilitas mengamati nilai tunggal sama dengan 0 karena jumlah nilai yang dapat diasumsikan oleh variabel acak tidak terbatas.Misalnya, variabel acak X dapat mengambil semua nilai selama interval bilangan real.Kemudian probabilitas bahwa X berada dalam himpunan hasil A, P (A), didefinisikan sebagai area above A dan under a curve.Kurva, yang merepresentasikan fungsi p (x), harus memenuhi yang berikut:

* The curve has no negative values (p(x) > 0 for all x)
* The total area under the curve is equal to 1.

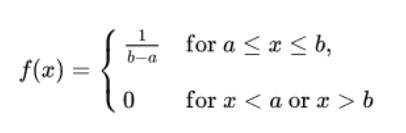
Kurva yang memenuhi persyaratan ini sering dikenal sebagai density curve.Beberapa contoh distribusi probabilitas kontinu adalah normal distribution, exponential distribution, beta distribution, dll.

### Note :

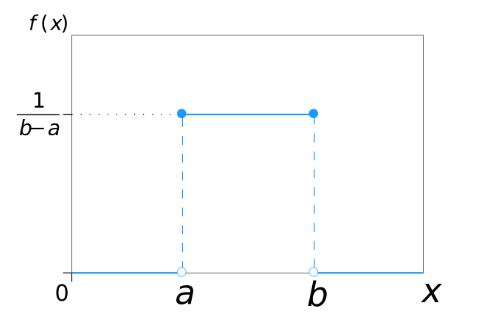
Ada jenis distribusi lain yang sering muncul : cumulative distribution function. Semua variabel acak (diskrit dan kontinu) memiliki cumulative distribution function. Ini adalah fungsi yang memberikan probabilitas bahwa variabel acak X kurang dari atau sama dengan x, untuk setiap nilai x. Untuk variabel acak diskrit, fungsi distribusi kumuluatif ditemukan dengan menjumlahkan probabilitas.

# Uniform Distribution

Mungkin salah satu yang paling sederhana dan berguna. Probabiliy distribution function dari continous uniform distribution adalah

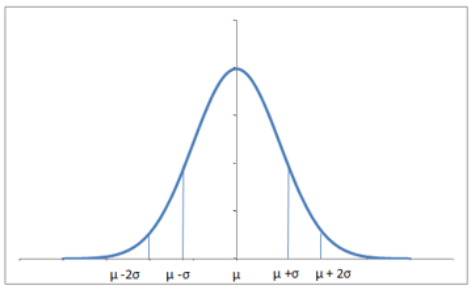


Karena setiap interval angka dengan lebar yang sama memiliki probabilitas yang sama untuk diamati, kurva yang menggambarkan distribusi tersebut adalah persegi panjang, dengan tinggi konstan melintasi interval dan 0 tinggi di tempat lain.Karena area di bawah kurva harus sama dengan 1, panjang interval menentukan tinggi kurva.Gambar berikut menunjukkan distribusi seragam dalam interval (a, b).Perhatikan karena luasnya harus 1. Tinggi diatur ke 1 / (b − a).



# Normal Distributon

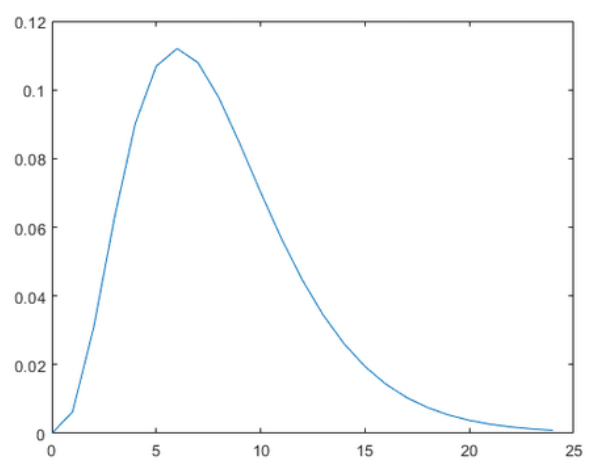
Memiliki kurva kerapatan berbentuk loncong yang dijelaskan dengan mean dan deviasi standar Kurva kepadatan simetris, berpusat di sekitar meannya, dengan penyebarannya ditentukan oleh deviasi standarnya yang menunjukkan bahwa data di dekat mean lebih sering terjadi daripada data yang jauh dari mean.



Hampir 68% data berada dalam jarak satu standard deviation dari mean di kedua sisi dan 95% dalam dua standard deviation. Juga perlu disebutkan bahwa distribusi dengan mean 0 dan standard deviation 1 disebut srandard normal distribution.

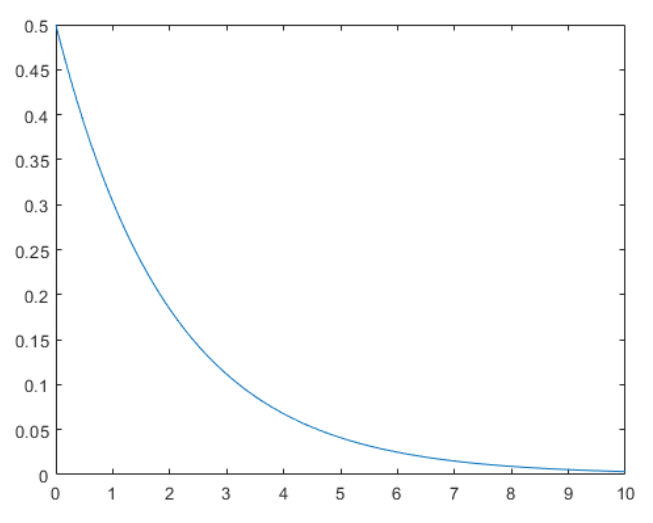
# Gamma Distribution

Two-parameter family dari continuous probability distributions. Jarang digunakan dalam bentuk mentahnya. Distribusi lain yang populer digunakan seperti exponential, chi-squared, erlang adalah kasus khusus dari distribusi gama.



# Exponential Distribution

Menggambarkan waktu antara peristiwa dalam Poisson point procces, yaitu proses di mana peristiwa terjadi terus menerus dan independen pada tingkat rata-rata yang konstan.



# Poisson Distribution

Digunakan untuk memodelkan beberapa kali suatu peristiwa terjadi dalam interval waktu. Misalnya, jumlah pengguna yang berkunjung di situs web dalam satu interval dapat dianggap sebagai proses Poisson. Distribusi poisson dijelaskan dalam tingat () di mana peristiwa terjadi. Suatu peristiwa dapat terjadi 0, 1, 2, … kali dalam satu interval. Jumlah rata-rata kejadian dalam suatu interval dintentukan lamda.

# Binomial Distribution

Distribusi dimana hasil yang mingkin hanya dua, sukses atau gagal, untung atau rugi, menang atau kalah, dimana probabilitas keberhasilan dan kegagalan sama untuk semua percobaan. Namun, hasil tidak harus sama kemungkinannya, dan setiap percobaan tidak bergantung satu sama lain. Parameter dari distribusi binomial adalah n dan p, dimana n adalah jumlah total percobaan dan p adalah probabilitas keberhasilan pada setiap percobaan.

# Bernoilli Distribution

Hanya memiliki dua kemungkinan hasil, yaitu 1 (berhasil) dan 0 (gagal), dan dalam sekali percobaan, misalnya lemparan koin. Jadi variabel random X yang berdistribusi Bernoulli dapat mengambil nilai 1 dengan probabilitas berhasil, p, dan bernilai 0 dengan probabilitas gagal, 1 atau 1-p. Probabilitas keberhasilan dan kegagalan tidak habur sama besarnya. Bernoulli distribution adalah kasus khusu dari binomial distribution di mana percobaan dilakukan tunggal (n=1)

# Confidence Interval

Jenis estimasi yang dihitung dari data statistik yang diamati. CI mengukur seberapa akurat mean sebuah sample mewakili (mencakup) nilai Mean Populasi sesungguhnya. Jadi secara singkat, CI adalah rentang antara dua nilai di mana nilai suatu Sample Mean tepat berada di tengah-tengahnya.

Bayangkan kita memutuskan untuk mengukur 40 pria secara acak di kota kita dan kita mendapatkan sampel tinggi rata-rata x̅ = 175 cm.Kita mungkin mendekati tinggi asli populasi (μ), tetapi kemungkinan besar nilai sebenarnya berada di antara 170 cm dan 180 cm.Paling akurat untuk mengatakan bahwa tinggi rata-rata pria di kota kita berada di antara interval tertentu [170 cm,180 cm].

**Why do we want a range?Because in real life, we are concerned about the confidence of our estimates.**

Biasanya ketika kita meminta seseorang untuk menebak tinggi orang di kota tertentu, orang tersebut lebih cenderung mengatakan sesuatu seperti: "Saya yakin tinggi orang-orang di Semarang antara 175-180cm" daripada menyebutkan angka tunggal seperti "Tinggi orang-orang di kota Semarang adalah 175cm”.

Kita sebagai manusia juga suka melampirkan tingkat kepercayaan saat kita memberikan perkiraan.Seperti - "Saya 90% yakin bahwa...".

**A confidence interval is a range within which you expect the population parameter to be.**

# How to Calculate Confidence Interval

Perhitungan confidence interval melibatkan estimasi terbaik yang diperoleh sample dan margin of error. Margin of error is z \* Estimated SE, SE is standard error.

Bayangkan seorang pembuat batu bata khawatir apakah massa batu bata yang diproduksinya sesuai dengan spesifikasi. Dia telah mengukur massa rata-rata sampel 100 batu bata sama dengan 3 kg. Dia juga menemukan interval kepercayaan 95% antara 2,85 kg dan 3,15 kg. Artinya, dia dapat yakin 95% bahwa massa rata-rata semua batu bata yang dia produksi akan berada di antara 2,85 kg dan 3,15 kg.

# Calculation of CI Mean

Berdasarkan central limit theorem, pengambilan sampel dalam jumlah yang cukup dengan ukuran yang memadai akan menghasilkan distribusi sample means yang normal.

# Hypothesis Testing

Anggapan dasar/jawaban sementara terhadap masalah yang masih bersifat praduga karena masih harus dibuktikan kebenarannya. Hipotesis harus dapat diuji, baik dengan eksperimen atau obeservasi.

1. A new medicine you think might work.
2. A way of teaching you think might be better.
3. A possible location of new species.
4. A fairer way to administer standardized tests.

What is a Hypothesis Statement?

Jika kita akan mengajukan hipotesis, biasanya pernyataan dibuat.Pernyataan kita akan terlihat seperti ini: “Jika saya… (melakukan A pada variabel independen)….Kemudian (B akan terjadi pada variabel dependen).”Sebagai contoh:

1. If I (decrease the amount of water given to herbs) then (the herbs will increase in size).
2. If I (give patients counseling in addition to medication) then (their overall depression scale will decrease).
3. If I (give exams at noon instead of 7) then (student test scores will improve).
4. If I (look in this certain location) then (I am more likely to find new species).

Pernyataan hipotesis yang baik harus:

1. Include an “if” and “then” statement (according to the University of California).
2. Include both the independent and dependent variables.
3. Be testable by experiment, survey or other scientifically sound technique.
4. Be based on information in prior research (either yours or someone else’s).
5. Have design criteria (for engineering or programming projects).

# What is Hypothesis Testing

Hypothesis testing dalam statistik adalah cara kita menguji hasil survei atau eksperimen untuk melihat apakah kita memiliki hasil yang bermakna.Kita pada dasarnya menguji apakah hasil kita valid dengan mencari tahu kemungkinan bahwa hasil kita terjadi secara kebetulan.Jika hasil kita mungkin terjadi secara kebetulan, eksperimen tidak akan dapat diulang sehingga hanya sedikit berguna.

Pengujian hipotesis dapat menjadi salah satu aspek yang paling membingungkan bagi siswa, terutama karena sebelum kita dapat melakukan pengujian, kita harus mengetahui apa null hypothesis kita.Seringkali, masalah kata rumit yang kita hadapi sulit untuk diuraikan.Tapi itu lebih mudah dari yang kita pikirkan;yang perlu kita lakukan adalah:

1. Figure out your null hypothesis,
2. State your null hypothesis,
3. Choose what kind of test you need to perform,
4. Either support or reject the null hypothesis.

# What is the Null Hypothesis?

Null hypothesis selalu merupakan fakta yang diterima.Contoh sederhana null hypothesis yang secara umum diterima sebagai benar adalah:

1. DNA is shaped like a double helix.
2. There are 8 planets in the solar system (excluding Pluto).
3. Taking Vioxx can increase your risk of heart problems (a drug now taken off the market).

Hypothesis Testing Examples: Basic Example

Researcher berpendapat bahwa jika pasien operasi lutut menjalani terapi fisik dua kali seminggu (bukan 3 kali), masa pemulihannya akan lebih lama.Waktu pemulihan rata-rata untuk pasien operasi lutut adalah 8.2 minggu.

Hypothesis statement dalam pertanyaan ini adalah bahwa researcher percaya waktu pemulihan rata-rata lebih dari 8.2 minggu.Ini dapat ditulis dalam istilah matematika sebagai: H1: μ> 8.2

Selanjutnya, kita harus menyatakan null hypothesis, yang akan terjadi jika researcher salah.Pada contoh di atas, jika peneliti salah maka waktu pemulihan kurang dari atau sama dengan 8,2 minggu.Dalam matematika, yaitu: H0 μ ≤ 8.2

# Statistical Test Interpretation

Hasil uji hipotesis statistik harus diinterpretasikan agar kita dapat mulai membuat klaim.

Ada dua bentuk umum yang dapat diambil oleh statistical hypothesis test, dan harus ditafsirkan dengan cara yang berbeda.Dua bentuk tersebut adalah adalah p-value and critical values.

Interpret the p-value

Kita menggambarkan temuan sebagai statistically significant dengan menafsirkan p-value.

Statistical hypothesis test dapat mengembalikan nilai yang disebut p atau p-value.Ini adalah kuantitas yang dapat kita gunakan untuk menafsirkan atau mengukur hasil tes: reject or fail to reject the null hypothesis.Ini dilakukan dengan membandingkan p-value dengan nilai ambang yang dipilih sebelumnya yang disebut significance level.

Significance level sering kali disebut dengan Greek lower case letter - alpha.

Nilai umum yang digunakan untuk alpha adalah 5% atau 0,05.Alpha value yang lebih kecil menunjukkan interpretasi yang lebih kuat dari null hypothesis, seperti 1% atau 0,1%.

p-value dibandingkan dengan alpha value yang dipilih sebelumnya.Hasil signifikan secara statistik saat p-value is less than alpha.Ini menandakan bahwa: default hypothesis dapat ditolak.

1. If p-value > alpha: Fail to reject the null hypothesis (i.e. not signifiant result).
2. If p-value <= alpha: Reject the null hypothesis (i.e. significant result).

For example, if we were performing a test of whether a data sample was normal and we calculated a p-value of .07, we could state something like:

The test found that the data sample was normal, failing to reject the null hypothesis at a 5% significance level.