**Lebih lanjut tentang integritas dan kepatuhan data**

Bacaan ini menggambarkan pentingnya integritas data dengan menggunakan contoh data perusahaan global. Definisi istilah yang relevan dengan integritas data akan diberikan di bagian akhir.

**Skenario: tanggal kalender untuk perusahaan global**

Tanggal kalender direpresentasikan dalam berbagai bentuk pendek yang berbeda. Tergantung di mana Anda tinggal, format yang berbeda mungkin digunakan.

* Di beberapa negara, **12/10/20** (DD/MM/YY) berarti 12 Oktober 2020.
* Di negara lain, standar nasionalnya adalah YYYY-MM-DD sehingga 12 Oktober 2020 menjadi **12-10-2020** .
* Di Amerika Serikat, (MM/DD/YY) adalah format yang diterima sehingga 12 Oktober 2020 akan menjadi **10/12/20** .

Sekarang, pikirkan apa yang akan terjadi jika Anda bekerja sebagai analis data untuk perusahaan global dan tidak memeriksa format tanggal. Ya, integritas data Anda mungkin dipertanyakan. Analisis data apa pun akan menjadi tidak akurat. Bayangkan memesan inventaris tambahan untuk bulan Desember padahal sebenarnya dibutuhkan di bulan Oktober!

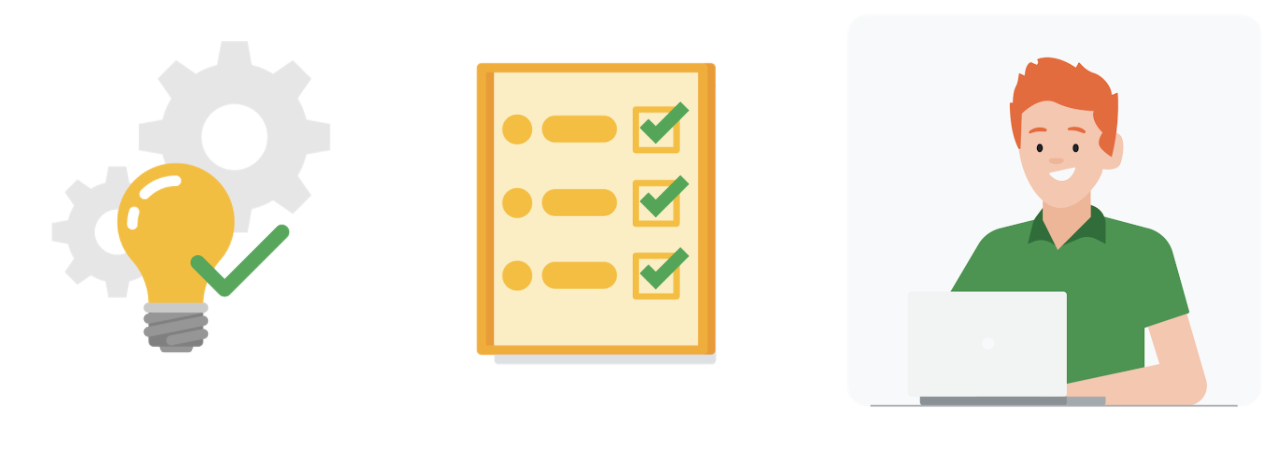
Analisis yang baik bergantung pada integritas data, dan integritas data biasanya bergantung pada penggunaan format umum. Jadi penting untuk memeriksa ulang bagaimana tanggal diformat untuk memastikan apa yang Anda anggap 10 Desember 2020 sebenarnya bukan 12 Oktober 2020, dan sebaliknya.

Berikut beberapa hal lain yang harus diperhatikan:

* **Replikasi data membahayakan integritas data:** Melanjutkan contoh ini, bayangkan Anda meminta rekanan internasional Anda untuk memverifikasi tanggal dan tetap menggunakan satu format. Seorang analis menyalin kumpulan data besar untuk memeriksa tanggalnya. Namun karena masalah memori, hanya sebagian dari kumpulan data yang disalin. Analis akan memverifikasi dan menstandarisasi data yang tidak lengkap. Sebagian kumpulan data tersebut akan disertifikasi sebagai patuh, namun kumpulan data lengkap masih berisi tanggal yang belum diverifikasi. Dua versi kumpulan data dapat menimbulkan hasil yang tidak konsisten. Audit akhir atas hasil akan sangat penting untuk mengungkap apa yang terjadi dan mengoreksi seluruh tanggal.
* **Transfer data membahayakan integritas data:** Analis lain memeriksa tanggal di spreadsheet dan memilih untuk mengimpor data yang divalidasi dan distandarisasi kembali ke database. Namun misalkan kolom tanggal dari spreadsheet salah diklasifikasikan sebagai kolom teks selama proses impor (transfer) data. Sekarang beberapa tanggal dalam database disimpan sebagai string teks. Pada titik ini, data perlu dibersihkan untuk mengembalikan integritasnya.
* **Manipulasi data yang membahayakan integritas data:** Saat memeriksa tanggal, analis lain memperhatikan apa yang tampak sebagai catatan duplikat dalam database dan menghapusnya. Namun ternyata analis tersebut menghapus catatan unik untuk anak perusahaan suatu perusahaan dan bukan catatan duplikat untuk perusahaan tersebut. Kumpulan data Anda sekarang tidak memiliki data dan data harus dipulihkan agar lengkap.

**Kesimpulan**

Untungnya, dengan format tanggal standar dan kepatuhan semua orang dan sistem yang bekerja dengan data, integritas data dapat dipertahankan. Namun dari mana pun data Anda berasal, selalu pastikan untuk memeriksa apakah data tersebut valid, lengkap, dan bersih sebelum Anda memulai analisis apa pun.



**Referensi: Batasan data dan contohnya**

Seiring kemajuan dalam perjalanan data, Anda akan menemukan banyak jenis batasan data (atau kriteria yang menentukan validitas). Tabel di bawah ini memberikan definisi dan contoh istilah batasan data yang mungkin Anda temui.

| **Batasan data** | **Definisi** | **Contoh** |
| --- | --- | --- |
| **Tipe data** | Nilai harus berjenis tertentu: tanggal, angka, persentase, Boolean, dll. | Jika tipe datanya adalah tanggal, angka tunggal seperti 30 akan gagal dalam batasan dan menjadi tidak valid |
| **Rentang data** | Nilai harus berada di antara nilai maksimum dan minimum yang telah ditentukan sebelumnya | Jika rentang datanya 10-20, nilai 30 akan gagal memenuhi batasan dan menjadi tidak valid |
| **Wajib** | Nilai tidak boleh dibiarkan kosong atau kosong | Jika usia wajib diisi, nilai tersebut harus diisi |
| **Unik** | Nilai tidak boleh memiliki duplikat | Dua orang tidak dapat memiliki nomor ponsel yang sama dalam area layanan yang sama |
| **Pola ekspresi reguler (regex).** | Nilai harus sesuai dengan pola yang ditentukan | Nomor telepon harus cocok dengan ###-###-#### (karakter lain tidak diperbolehkan) |
| **Validasi lintas bidang** | Kondisi tertentu untuk beberapa bidang harus dipenuhi | Nilai adalah persentase dan nilai dari beberapa bidang harus berjumlah hingga 100% |
| **Kunci utama** | (Hanya basis data) nilai harus unik per kolom | Tabel database tidak boleh memiliki dua baris dengan nilai kunci utama yang sama. Kunci utama adalah pengidentifikasi dalam database yang mereferensikan kolom yang setiap nilainya unik. Informasi lebih lanjut tentang kunci primer dan kunci asing disediakan kemudian dalam program ini. |
| **Keanggotaan set** | (Hanya database) nilai untuk kolom harus berasal dari sekumpulan nilai diskrit | Nilai kolom harus disetel ke Ya, Tidak, atau Tidak Berlaku |
| **Kunci asing** | (Hanya database) nilai untuk kolom harus berupa nilai unik yang berasal dari kolom di tabel lain | Dalam database wajib pajak AS, kolom Negara Bagian harus berupa negara bagian atau teritori yang valid dengan kumpulan nilai yang dapat diterima yang ditentukan dalam tabel Negara Bagian terpisah |
| **Ketepatan** | Sejauh mana data sesuai dengan entitas sebenarnya yang diukur atau dijelaskan | Jika nilai kode pos divalidasi berdasarkan lokasi jalan, keakuratan data akan meningkat. |
| **Kelengkapan** | Sejauh mana data memuat semua komponen atau ukuran yang diinginkan | Jika data untuk profil pribadi memerlukan warna rambut dan warna mata, dan keduanya dikumpulkan, maka data sudah lengkap. |
| **Konsistensi** | Sejauh mana data dapat diulang dari berbagai titik masuk atau pengumpulan | Jika pelanggan memiliki alamat yang sama di database penjualan dan perbaikan, datanya konsisten. |

# Tujuan dan data yang selaras

Anda bisa mendapatkan wawasan yang kuat dan membuat kesimpulan yang akurat ketika data selaras dengan tujuan bisnis. Sebagai seorang analis data, keselarasan adalah sesuatu yang perlu Anda nilai. Penyelarasan yang baik berarti datanya relevan dan dapat membantu Anda memecahkan masalah bisnis atau menentukan tindakan untuk mencapai tujuan bisnis tertentu.

Dalam bacaan ini, Anda akan meninjau tujuan bisnis yang terkait dengan tiga skenario. Anda akan mempelajari bagaimana data yang bersih dan tujuan bisnis yang selaras dapat membantu Anda menghasilkan kesimpulan yang akurat. Selain itu, Anda akan mempelajari bagaimana variabel baru yang ditemukan selama analisis data dapat menyebabkan Anda menyiapkan batasan data sehingga Anda dapat menjaga data tetap selaras dengan tujuan bisnis.

## Data yang bersih + keselarasan dengan tujuan bisnis = kesimpulan yang akurat

### **Tujuan bisnis**

Manajer akun di Impress Me, sebuah layanan berlangganan konten online, ingin mengetahui seberapa cepat pengguna melihat konten setelah langganan mereka diaktifkan.



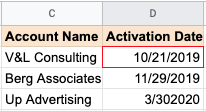
Untuk memulai, analis data memverifikasi bahwa data yang diekspor ke spreadsheet bersih dan mengonfirmasi bahwa data yang diperlukan (saat pengguna mengakses konten) tersedia. Mengetahui hal ini, analis memutuskan adanya keselarasan data yang baik dengan tujuan bisnis. Yang belum ada hanyalah mengetahui berapa lama waktu yang dibutuhkan setiap pengguna untuk melihat konten setelah langganan mereka diaktifkan.

Berikut langkah pemrosesan data yang dilakukan analis untuk pengguna dari akun bernama V&L Consulting. (Langkah-langkah ini akan diulangi untuk setiap akun berlangganan, dan untuk setiap pengguna yang terkait dengan akun tersebut.)

### **Langkah 1**

| **Langkah pemrosesan data** | **Sumber data** |
| --- | --- |
| Cari tanggal aktivasi V&L Consulting | Spreadsheet akun |

**Data yang relevan di spreadsheet:**

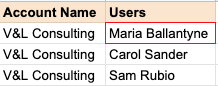


**Hasil** : 21 Oktober 2019

### **Langkah 2**

| **Langkah pemrosesan data** | **Sumber data** |
| --- | --- |
| Cari nama pengguna yang termasuk dalam akun V&L Consulting | Spreadsheet akun (tab pengguna) |

**Data yang relevan di spreadsheet** :

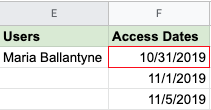


**Hasil** : Maria Ballantyne

### **Langkah 3**

| **Langkah pemrosesan data** | **Sumber data** |
| --- | --- |
| Temukan tanggal akses konten pertama untuk Maria B. | Spreadsheet penggunaan konten |

**Data yang relevan di spreadsheet:**

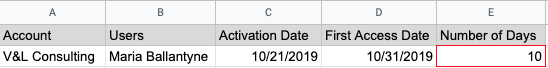


**Hasil** : 31 Oktober 2019

### **Langkah 4**

| **Langkah pemrosesan data** | **Sumber data** |
| --- | --- |
| Hitung waktu antara aktivasi dan penggunaan konten pertama untuk Maria B. | Perhitungan spreadsheet baru |

**Data yang relevan di spreadsheet** :



**Hasil** : 10 hari

### **Kiat profesional 1**

Dalam proses di atas, analis dapat menggunakan **VLOOKUP** untuk mencari data pada Langkah 1, 2, dan 3 untuk mengisi nilai dalam spreadsheet pada Langkah 4.[VLOOKUP](https://support.microsoft.com/en-us/office/vlookup-function-0bbc8083-26fe-4963-8ab8-93a18ad188a1)adalah fungsi spreadsheet yang mencari nilai tertentu dalam kolom untuk mengembalikan informasi terkait. Menggunakan VLOOKUP dapat menghemat banyak waktu; tanpanya, Anda harus mencari tanggal dan nama secara manual.

Mengacu kepada[VLOOKUP](https://support.google.com/docs/answer/3093318?hl=en)halaman di Pusat Bantuan Google untuk mengetahui cara menggunakan fungsi tersebut di Google Spreadsheet.

### **Kiat profesional 2**

Pada Langkah 4 dari proses di atas, analis dapat menggunakan fungsi **DATEDIF** untuk secara otomatis menghitung selisih antara tanggal di kolom C dan kolom D. Fungsi tersebut dapat menghitung jumlah hari antara dua tanggal. 

Lihat Dukungan Microsoft[TANGGALJIKA](https://support.microsoft.com/en-us/office/datedif-function-25dba1a4-2812-480b-84dd-8b32a451b35c)halaman untuk cara menggunakan fungsi di Excel. Itu[HARI360](https://support.microsoft.com/en-us/office/days360-function-b9a509fd-49ef-407e-94df-0cbda5718c2a)fungsi melakukan hal yang sama dalam spreadsheet akuntansi yang menggunakan tahun 360 hari (dua belas bulan 30 hari).

Mengacu kepada[TANGGALJIKA](https://support.google.com/docs/answer/6055612?hl=en)halaman di Pusat Bantuan Google untuk mengetahui cara menggunakan fungsi tersebut di Google Spreadsheet.

## Penyelarasan dengan tujuan bisnis + pembersihan data tambahan = kesimpulan yang akurat

### **Tujuan bisnis**

Cloud Gate, sebuah perusahaan perangkat lunak, baru-baru ini menyelenggarakan serangkaian webinar publik sebagai pengenalan produk gratis. Analis data dan manajer program webinar ingin mengidentifikasi perusahaan yang memiliki lima orang atau lebih yang menghadiri sesi ini. Mereka ingin memberikan daftar perusahaan ini kepada manajer penjualan yang dapat menindaklanjuti potensi penjualan.



Data kehadiran webinar mencakup kolom dan data yang ditunjukkan di bawah ini.

| **Nama** | **< Nama depan > < Nama belakang >** | **Ini adalah informasi wajib yang harus diserahkan oleh peserta** |
| --- | --- | --- |
| **Alamat email** | xxxxx@ perusahaan .com | Ini adalah informasi wajib yang harus diserahkan oleh peserta |
| **Perusahaan** | < Nama perusahaan > | Ini adalah informasi opsional yang dapat diberikan oleh peserta |

### **Pembersihan data**

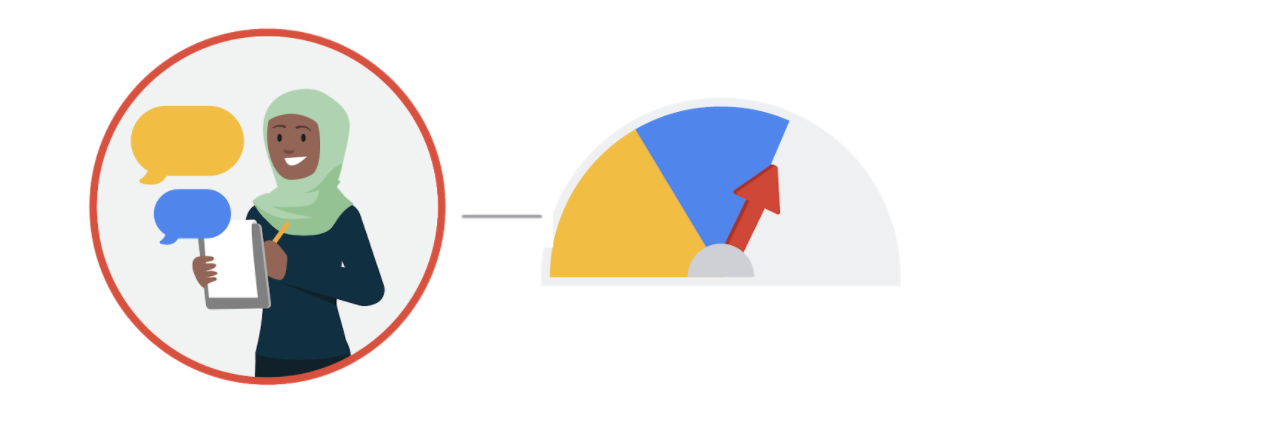
Data kehadiran webinar tampaknya sejalan dengan tujuan bisnis. Namun analis data dan manajer program memutuskan bahwa pembersihan data diperlukan sebelum analisis. Mereka menganggap pembersihan data diperlukan karena:

* Nama perusahaan bukanlah bidang wajib. Jika nama perusahaan kosong, mungkin ditemukan dari alamat email. Misalnya, jika alamat emailnya adalah nama pengguna@google.com, kolom perusahaan dapat diisi dengan Google untuk analisis data. Langkah pembersihan data ini mengasumsikan bahwa orang-orang dengan alamat email yang ditetapkan perusahaan menghadiri webinar untuk tujuan bisnis.
* Peserta dapat memasukkan nama apa saja. Karena kehadiran di serangkaian webinar sedang diperiksa, mereka perlu memvalidasi nama dengan alamat email unik. Misalnya, jika Joe Cox menghadiri dua webinar namun masuk sebagai Joe Cox untuk satu webinar dan Joseph Cox untuk webinar lainnya, dia akan dihitung sebagai dua orang yang berbeda. Untuk mencegah hal ini, mereka perlu memeriksa alamat email uniknya untuk menentukan bahwa dia adalah orang yang sama. Setelah validasi, Joseph Cox dapat diubah menjadi Joe Cox untuk mencocokkan instance lainnya.

## Penyelarasan dengan tujuan bisnis + variabel yang baru ditemukan + kendala = kesimpulan yang akurat

### **Tujuan bisnis**

Sebuah perusahaan bimbingan belajar sepulang sekolah, A+ Education, ingin mengetahui apakah ada jumlah minimum jam bimbingan belajar yang diperlukan sebelum siswa mengalami peningkatan setidaknya 10% dalam nilai penilaian mereka.



Analis data berpendapat ada keselarasan yang baik antara data yang tersedia dan tujuan bisnis karena:

* Siswa masuk dan keluar dari sistem untuk setiap sesi bimbingan belajar, dan jumlah jam dilacak
* Skor penilaian dicatat secara berkala

### **Batasan data untuk variabel baru**

Setelah melihat data, analis data menemukan bahwa ada variabel lain yang perlu dipertimbangkan. Beberapa siswa memiliki sesi mingguan yang konsisten sementara siswa lainnya menjadwalkan sesi secara lebih acak meskipun jumlah jam les mereka sama. Data tersebut tidak selaras dengan tujuan bisnis awal seperti yang diperkirakan sebelumnya, sehingga analis menambahkan batasan data untuk fokus hanya pada siswa dengan sesi mingguan yang konsisten. Modifikasi ini membantu mendapatkan gambaran yang lebih akurat tentang waktu pendaftaran yang diperlukan untuk mencapai peningkatan skor penilaian sebesar 10%.

## Poin-poin penting

Semoga contoh-contoh ini memberi Anda gambaran tentang apa yang harus dicari untuk mengetahui apakah data Anda selaras dengan tujuan bisnis Anda.

* Jika terdapat data yang bersih dan keselarasan yang baik, Anda bisa mendapatkan wawasan yang akurat dan membuat kesimpulan yang didukung data tersebut.
* Jika ada penyelarasan yang baik tetapi datanya perlu dibersihkan, bersihkan data tersebut sebelum Anda melakukan analisis.
* Jika data hanya selaras sebagian dengan tujuan, pikirkan bagaimana Anda dapat mengubah tujuan, atau gunakan batasan data untuk memastikan bahwa subkumpulan data lebih selaras dengan tujuan bisnis.

# Apa yang harus dilakukan ketika Anda menemukan masalah dengan data Anda

Saat Anda bersiap untuk analisis data, Anda mungkin menyadari bahwa Anda tidak memiliki data yang diperlukan atau tidak memiliki cukup data. Dalam beberapa kasus, Anda dapat menggunakan apa yang disebut sebagai data proxy sebagai pengganti data sebenarnya. Anggap saja seperti mengganti mentega dengan minyak dalam resep saat Anda tidak memiliki mentega. Dalam kasus lain, tidak ada pengganti yang masuk akal dan satu-satunya pilihan Anda adalah mengumpulkan lebih banyak data.

Pertimbangkan masalah data berikut dan saran tentang cara mengatasinya.

## Masalah data 1: tidak ada data

| **Solusi yang memungkinkan** | **Contoh solusi dalam kehidupan nyata** |
| --- | --- |
| Kumpulkan data dalam skala kecil untuk melakukan analisis awal dan kemudian minta waktu tambahan untuk menyelesaikan analisis setelah Anda mengumpulkan lebih banyak data. | Jika Anda mensurvei karyawan tentang pendapat mereka tentang kinerja baru dan rencana bonus, gunakan sampel untuk analisis awal. Lalu, minta waktu 3 minggu lagi untuk mengumpulkan data seluruh karyawan. |
| Jika tidak ada waktu untuk mengumpulkan data, lakukan analisis menggunakan data proksi dari kumpulan data lain.  Ini adalah solusi paling umum. | Jika Anda menganalisis waktu perjalanan puncak bagi penumpang namun tidak memiliki data untuk kota tertentu, gunakan data dari kota lain dengan ukuran dan demografi yang serupa. |

## Masalah data 2: data terlalu sedikit

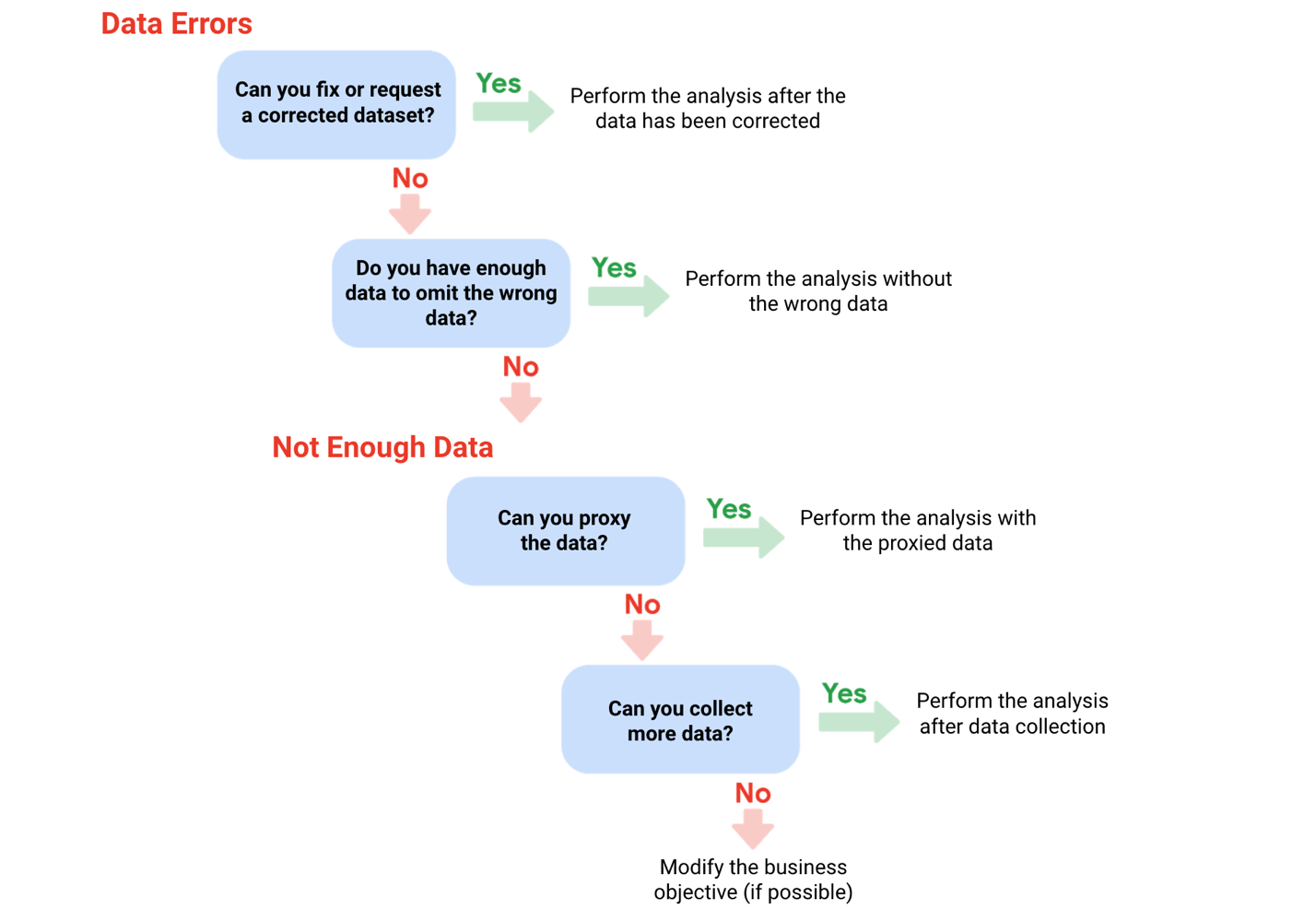
| **Solusi yang memungkinkan** | **Contoh solusi dalam kehidupan nyata** |
| --- | --- |
| Lakukan analisa dengan menggunakan data proksi beserta data sebenarnya. | Jika Anda menganalisis tren pemilik anjing Golden Retriever, perbesar kumpulan data Anda dengan menyertakan data dari pemilik anjing labrador. |
| Sesuaikan analisis Anda agar selaras dengan data yang sudah Anda miliki. | Jika Anda tidak memiliki data untuk kelompok usia 18 hingga 24 tahun, lakukan analisis namun perhatikan batasan berikut dalam laporan Anda: kesimpulan ini hanya berlaku untuk orang dewasa berusia 25 tahun ke atas . |

## Masalah data 3: data salah, termasuk data yang error\*

| **Solusi yang memungkinkan** | **Contoh solusi dalam kehidupan nyata** |
| --- | --- |
| Jika Anda memiliki data yang salah karena persyaratannya disalahpahami, komunikasikan kembali persyaratan tersebut. | Jika Anda membutuhkan data pemilih perempuan dan menerima data pemilih laki-laki, nyatakan kembali kebutuhan Anda. |
| Identifikasi kesalahan dalam data dan, jika mungkin, perbaiki pada sumbernya dengan mencari pola kesalahan tersebut. | Jika data Anda ada dalam spreadsheet dan terdapat pernyataan kondisional atau boolean yang menyebabkan penghitungan salah, ubah pernyataan kondisional alih-alih hanya memperbaiki nilai terhitung. |
| Jika Anda tidak dapat memperbaiki sendiri kesalahan data, Anda dapat mengabaikan data yang salah dan melanjutkan analisis jika ukuran sampel Anda masih cukup besar dan mengabaikan data tidak akan menyebabkan bias sistematis. | Jika kumpulan data Anda diterjemahkan dari bahasa lain dan beberapa terjemahannya tidak masuk akal, abaikan data dengan terjemahan yang buruk dan lanjutkan analisis data lainnya. |

\*Catatan penting: terkadang data yang error bisa menjadi tanda peringatan bahwa data tersebut tidak dapat diandalkan. Gunakan penilaian terbaik Anda.

### Gunakan pohon keputusan berikut sebagai pengingat tentang cara menangani kesalahan data atau tidak cukup data:

1. Bisakah Anda memperbaiki atau meminta kumpulan data yang diperbaiki? TIDAK 2. Apakah Anda mempunyai cukup data untuk menghilangkan data yang salah? TIDAK 3. Bisakah Anda mem-proxy datanya? TIDAK 4. Bisakah Anda mengumpulkan lebih banyak data? TIDAK Ubah tujuan bisnis (jika memungkinkan)

# Menghitung ukuran sampel

Sebelum Anda menggali lebih dalam tentang ukuran sampel, kenali istilah dan definisi berikut:

| **Terminologi** | **Definisi** |
| --- | --- |
| **Populasi** | Seluruh kelompok yang Anda minati untuk studi Anda. Misalnya, jika Anda mensurvei orang-orang di perusahaan Anda, populasinya adalah seluruh karyawan di perusahaan Anda. |
| **Sampel** | Sebagian dari populasi Anda. Sama seperti sampel makanan, disebut sampel karena hanya sekedar rasa. Jadi, jika perusahaan Anda terlalu besar untuk mensurvei setiap individu, Anda dapat mensurvei sampel yang mewakili populasi Anda. |
| **Margin kesalahan** | Karena sampel digunakan untuk mewakili suatu populasi, hasil sampel diharapkan berbeda dari hasil yang akan diperoleh jika Anda mensurvei seluruh populasi. Perbedaan ini disebut margin of error. Semakin kecil margin kesalahannya, semakin dekat hasil sampel dengan hasil yang akan diperoleh jika Anda mensurvei seluruh populasi. |
| **Tingkat kepercayaan diri** | Seberapa yakin Anda dengan hasil survei. Misalnya, tingkat kepercayaan 95% berarti jika Anda menjalankan survei yang sama sebanyak 100 kali, Anda akan mendapatkan hasil serupa sebanyak 95 kali dari 100 kali survei tersebut. Tingkat kepercayaan diri ditargetkan sebelum Anda memulai studi karena akan mempengaruhi seberapa besar margin of error Anda di akhir studi. |
| **Interval kepercayaan** | Kisaran nilai yang mungkin diperoleh dari hasil populasi pada tingkat kepercayaan penelitian. Kisaran ini adalah hasil sampel +/- margin kesalahan. |
| **Signifikansi statistik** | Penentuan apakah hasil Anda bisa jadi karena kebetulan atau tidak. Semakin besar signifikansinya, semakin kecil kemungkinannya. |

## Hal yang perlu diingat saat menentukan ukuran sampel Anda

Saat menentukan ukuran sampel, berikut beberapa hal yang perlu diingat:

* Jangan gunakan ukuran sampel yang kurang dari 30. Telah terbukti secara statistik bahwa 30 adalah ukuran sampel terkecil dimana hasil rata-rata suatu sampel mulai mewakili hasil rata-rata suatu populasi.
* Tingkat kepercayaan yang paling umum digunakan adalah 95%, namun 90% dapat digunakan dalam beberapa kasus.

Tingkatkan ukuran sampel untuk memenuhi kebutuhan spesifik proyek Anda:

* Untuk tingkat kepercayaan **yang lebih tinggi** , gunakan ukuran sampel yang lebih besar
* Untuk **mengurangi** margin kesalahan, gunakan ukuran sampel yang lebih besar
* Untuk signifikansi statistik **yang lebih besar** , gunakan ukuran sampel yang lebih besar

**Catatan:** Kalkulator ukuran sampel menggunakan rumus statistik untuk menentukan ukuran sampel. Informasi lebih lanjut mengenai hal ini akan dibahas dalam kursus ini! Pantau terus.

### **Mengapa sampel minimal 30?**

Rekomendasi ini didasarkan pada **Central Limit Theorem (CLT)** di bidang probabilitas dan statistik. Seiring bertambahnya ukuran sampel, hasilnya semakin mirip dengan distribusi normal (berbentuk lonceng) dari sejumlah besar sampel. Sampel sebanyak 30 merupakan ukuran sampel terkecil yang CLT-nya masih valid. Peneliti yang mengandalkan **analisis regresi** – metode statistik untuk menentukan hubungan antara variabel terkontrol dan variabel dependen – juga lebih memilih sampel minimal 30 orang.

Masih penasaran? Tanpa terlalu mendalami matematika, lihat artikel ini:

* [Teorema Limit Pusat (CLT)](https://www.investopedia.com/terms/c/central_limit_theorem.asp): Artikel dari Investopedia ini menjelaskan Teorema Batas Pusat dan menjelaskan secara singkat bagaimana penerapannya pada analisis indeks saham.
* [Rumus Ukuran Sampel](https://www.statisticssolutions.com/dissertation-resources/sample-size-calculation-and-sample-size-justification/sample-size-formula/): Artikel dari Statistics Solutions ini memberikan sedikit lebih banyak detail tentang mengapa beberapa peneliti menggunakan 30 sebagai ukuran sampel minimum.

## Ukuran sampel bervariasi berdasarkan masalah bisnis

Ukuran sampel akan bervariasi berdasarkan jenis masalah bisnis yang Anda coba selesaikan.

Misalnya, jika Anda tinggal di kota dengan populasi 200.000 jiwa dan terdapat 180.000 orang yang merespons survei, itu adalah ukuran sampel yang besar. Namun tanpa benar-benar melakukan hal tersebut, seperti apa ukuran sampel yang lebih kecil dan dapat diterima?

Apakah 200 orang akan baik-baik saja jika orang yang disurvei mewakili setiap distrik di kota tersebut?

**Jawaban** : Tergantung taruhannya.

* Ukuran sampel sebanyak 200 mungkin cukup besar jika masalah bisnis Anda adalah mengetahui bagaimana perasaan penghuni terhadap perpustakaan baru
* Ukuran sampel sebesar 200 mungkin tidak cukup besar jika masalah bisnis Anda adalah menentukan bagaimana warga akan memilih untuk mendanai perpustakaan

Anda mungkin dapat menerima margin kesalahan yang lebih besar dalam mensurvei bagaimana perasaan penghuni terhadap perpustakaan baru dibandingkan dengan mensurvei penduduk tentang bagaimana mereka akan memilih untuk mendanai perpustakaan tersebut. Oleh karena itu, kemungkinan besar Anda akan menggunakan ukuran sampel yang lebih besar untuk survei pemilih.



## Ukuran sampel yang lebih besar memiliki biaya yang lebih tinggi

Anda juga harus mempertimbangkan biaya dibandingkan manfaat dari hasil yang lebih akurat dengan ukuran sampel yang lebih besar. Seseorang yang mencoba memahami preferensi konsumen terhadap produk baru tidak memerlukan ukuran sampel yang besar seperti seseorang yang mencoba memahami efek obat baru. Untuk keamanan obat, manfaatnya lebih besar daripada biaya menggunakan ukuran sampel yang lebih besar. Namun untuk preferensi konsumen, ukuran sampel yang lebih kecil dengan biaya yang lebih rendah dapat memberikan hasil yang cukup baik.



## Mengetahui dasar-dasarnya sangat membantu

Mengetahui dasar-dasarnya akan membantu Anda membuat pilihan yang tepat terkait ukuran sampel. Anda selalu dapat menyampaikan kekhawatiran jika menemukan ukuran sampel yang terlalu kecil. Kalkulator ukuran sampel juga merupakan alat yang hebat untuk ini. Kalkulator ukuran sampel memungkinkan Anda memasukkan tingkat kepercayaan dan margin kesalahan yang diinginkan untuk ukuran populasi tertentu. Mereka kemudian menghitung ukuran sampel yang diperlukan untuk mencapai hasil tersebut secara statistik.

Mengacu kepada[Tentukan Ukuran Sampel Terbaik](https://www.coursera.org/learn/process-data/lecture/mSj5A/determine-the-best-sample-size)video untuk demonstrasi kalkulator ukuran sampel, atau lihat[Kalkulator Ukuran Sampel](https://www.coursera.org/learn/process-data/supplement/ZqcDw/sample-size-calculator)membaca untuk informasi tambahan.



# Apa yang harus dilakukan jika tidak ada data

Sebelumnya, Anda telah mempelajari bagaimana Anda masih dapat melakukan analisis menggunakan data proxy jika Anda tidak memiliki data. Anda mungkin memiliki beberapa pertanyaan tentang data proksi, jadi bacaan ini akan memberi Anda beberapa contoh jenis kumpulan data yang dapat berfungsi sebagai sumber data alternatif.

## Contoh data proxy

Terkadang data untuk mendukung tujuan bisnis tidak tersedia. Di sinilah data proxy berguna. Lihatlah skenario berikut dan di mana data proxy masuk untuk setiap contoh:

| **Skenario bisnis** | **Bagaimana data proxy dapat digunakan** |
| --- | --- |
| Model mobil baru baru saja diluncurkan beberapa hari yang lalu dan dealer mobil tidak sabar menunggu hingga akhir bulan untuk mendapatkan data penjualan. Mereka menginginkan proyeksi penjualan sekarang. | Analis memproksikan jumlah klik terhadap spesifikasi mobil di website dealer sebagai perkiraan potensi penjualan di dealer tersebut. |
| Produk daging nabati baru baru saja tersedia di toko kelontong dan pemasok perlu memperkirakan permintaan selama empat tahun ke depan. | Analis memproksi data penjualan pengganti kalkun yang terbuat dari tahu yang telah ada di pasaran selama beberapa tahun. |
| Kamar Dagang ingin mengetahui bagaimana kampanye pariwisata akan berdampak pada perjalanan ke kota mereka, namun hasil dari kampanye tersebut belum tersedia untuk umum. | Analis tersebut memproksi data historis pemesanan maskapai penerbangan ke kota tersebut satu hingga tiga bulan setelah kampanye serupa dijalankan enam bulan sebelumnya. |

## Kumpulan data terbuka (publik).

Jika Anda adalah bagian dari organisasi besar, Anda mungkin memiliki akses ke banyak sumber data. Namun jika Anda mencari sesuatu yang spesifik atau sedikit di luar bidang bisnis Anda, Anda juga dapat menggunakan kumpulan data terbuka atau publik. (Anda dapat merujuk ke ini[Artikel sedang](https://medium.com/thinkdata/is-there-a-difference-between-open-data-and-public-data-2b8d5608b2f1)untuk penjelasan singkat tentang perbedaan antara data terbuka dan publik.)

Berikut ini contohnya. Vaksin versi hidung baru-baru ini tersedia. Sebuah klinik ingin mengetahui apa yang diharapkan dari kontraindikasi, namun baru saja mulai mengumpulkan data pihak pertama dari pasiennya. Kontraindikasi adalah suatu kondisi yang dapat menyebabkan pasien tidak menerima vaksin karena dampak buruknya jika diberikan **.** Untuk memperkirakan jumlah kemungkinan kontraindikasi, analis data memproksi kumpulan data terbuka dari uji coba vaksin versi injeksi. Analis memilih subset data dengan profil pasien yang paling cocok dengan susunan pasien di klinik.

Ada banyak cara untuk berbagi dan berkolaborasi pada data dalam komunitas. Kaggle ([kaggle.com](https://www.kaggle.com/)) yang kami perkenalkan sebelumnya, memiliki kumpulan data dalam berbagai format termasuk tipe paling dasar, file Comma Separated Values ​​(CSV).



### **Kumpulan data CSV, JSON, SQLite, dan BigQuery**

* CSV: Lihat ini[Pelanggan kartu kredit](https://www.kaggle.com/sakshigoyal7/credit-card-customers)kumpulan data, yang berisi informasi dari 10.000 pelanggan termasuk usia, gaji, status perkawinan, batas kartu kredit, kategori kartu kredit, dll. (CC0: Domain Publik, Sakshi Goyal).
* JSON: Lihat kumpulan data JSON ini[video YouTube yang sedang tren](https://www.kaggle.com/datasnaek/youtube-new)(CC0: Domain Publik, Mitchell J).
* SQLite: Lihat kumpulan data SQLite ini selama 24 tahun[Data kebakaran hutan di AS](https://www.kaggle.com/rtatman/188-million-us-wildfires)(CC0: Domain Publik, Rachael Tatman).
* BigQuery: Lihat ini[Google Analitik 360](https://www.kaggle.com/bigquery/google-analytics-sample)contoh kumpulan data dari Google Merchandise Store (CC0 Public Domain, Google BigQuery).

Lihat Kaggle[dokumentasi untuk kumpulan data](https://www.kaggle.com/docs/datasets)untuk informasi lebih lanjut dan cari serta jelajahi kumpulan data Anda sendiri di[kaggle.com/datasets](https://www.kaggle.com/datasets).

Seperti semua jenis kumpulan data lainnya, waspadai data duplikat dan 'Null' di kumpulan data terbuka. Null paling sering berarti bahwa bidang data tidak ditetapkan (dibiarkan kosong), namun terkadang Null dapat diartikan sebagai nilai, 0. Penting untuk memahami bagaimana Null digunakan sebelum Anda mulai menganalisis kumpulan data dengan data Null.

# Kalkulator ukuran sampel

Dalam bacaan ini, Anda akan mempelajari dasar-dasar kalkulator ukuran sampel, cara menggunakannya, dan cara memahami hasilnya. Kalkulator **ukuran sampel** memberi tahu Anda berapa banyak orang yang perlu Anda wawancarai (atau hal-hal yang perlu Anda uji) untuk mendapatkan hasil yang mewakili populasi target. Mari kita tinjau beberapa istilah yang akan Anda temui saat menggunakan kalkulator ukuran sampel:

* **Tingkat kepercayaan** : Kemungkinan bahwa ukuran sampel Anda secara akurat mencerminkan populasi yang lebih besar.
* **Margin of error** : Jumlah maksimum hasil sampel yang diharapkan berbeda dari populasi sebenarnya.
* **Populasi** : Ini adalah jumlah total yang Anda harapkan untuk diambil sampelnya.
* **Sampel** : Bagian dari suatu populasi yang mewakili populasi tersebut.
* **Perkiraan tingkat respons** : Jika Anda menjalankan survei terhadap individu, ini adalah persentase orang yang Anda harapkan akan menyelesaikan survei Anda dari mereka yang menerima survei.

## Cara menggunakan kalkulator ukuran sampel

Untuk menggunakan kalkulator ukuran sampel, Anda harus sudah menentukan ukuran populasi, tingkat kepercayaan, dan margin kesalahan yang dapat diterima sehingga Anda dapat memasukkannya ke dalam alat. Jika informasi ini sudah siap, lihat kalkulator ukuran sampel di bawah ini:

* [Kalkulator ukuran sampel oleh surveymonkey.com](https://www.surveymonkey.com/mp/sample-size-calculator/)
* [Kalkulator ukuran sampel oleh raosoft.com](http://www.raosoft.com/samplesize.html)

## Apa yang harus dilakukan dengan hasilnya

Setelah Anda memasukkan informasi Anda ke salah satu kalkulator ini, kalkulator ini akan memberi Anda ukuran sampel yang disarankan. Perlu diingat, ukuran sampel yang dihitung adalah jumlah **minimum** untuk mencapai apa yang Anda masukkan untuk tingkat kepercayaan dan margin kesalahan. Jika Anda mengerjakan survei, Anda juga perlu memikirkan perkiraan tingkat respons untuk mengetahui berapa banyak survei yang perlu Anda kirimkan. Misalnya, jika Anda memerlukan ukuran sampel 100 individu dan perkiraan tingkat respons Anda adalah 10%, Anda perlu mengirimkan survei Anda ke 1.000 individu untuk mendapatkan 100 respons yang Anda perlukan untuk analisis Anda.

Sekarang setelah Anda memiliki dasar-dasarnya, cobalah beberapa penghitungan menggunakan kalkulator ukuran sampel dan lihat kembali bacaan ini jika Anda memerlukan penyegaran tentang definisinya.

**Semua tentang margin kesalahan**

**Margin of error** adalah jumlah maksimum hasil sampel yang diharapkan berbeda dari populasi sebenarnya. Secara lebih teknis, margin kesalahan menentukan rentang nilai di bawah dan di atas hasil rata-rata sampel. Hasil rata-rata untuk seluruh populasi diharapkan berada dalam kisaran tersebut. Kita dapat lebih memahami margin of error dengan menggunakan beberapa contoh di bawah ini.

**Margin kesalahan dalam bisbol**



Bayangkan Anda sedang bermain bisbol dan Anda siap memukul. Penonton bersorak dan Anda bersiap untuk mencoba memukul bola. Pelempar mengirimkan bola cepat dengan kecepatan sekitar 90-95mph, yang membutuhkan waktu sekitar 400 milidetik (ms) untuk mencapai sarung tangan penangkap. Anda mengayun dan melewatkan lemparan pertama karena waktu Anda sedikit melenceng. Anda bertanya-tanya apakah Anda seharusnya mengayun sedikit lebih awal atau sedikit lebih lambat untuk melakukan home run. Perbedaan waktu tersebut dapat dianggap sebagai margin kesalahan, dan ini memberi tahu kita seberapa dekat atau jauh waktu Anda dari rata-rata ayunan home run.

**Margin kesalahan dalam pemasaran**

Margin of error juga penting dalam pemasaran. Mari kita gunakan pengujian A/B sebagai contoh. **Pengujian A/B** (atau pengujian terpisah) menguji dua variasi laman web yang sama untuk menentukan laman mana yang lebih berhasil dalam menarik lalu lintas pengguna dan menghasilkan pendapatan. Lalu lintas pengguna yang dimonetisasi dikenal sebagai **tingkat konversi** . Pengujian A/B memungkinkan pemasar menguji email, iklan, dan halaman arahan untuk menemukan data di balik apa yang berhasil dan apa yang tidak. Pemasar menggunakan **interval kepercayaan** (ditentukan oleh tingkat konversi dan margin kesalahan) untuk memahami hasilnya.

Misalnya, Anda melakukan pengujian A/B untuk membandingkan efektivitas dua baris subjek email yang berbeda untuk menarik orang agar membuka email. Anda menemukan bahwa baris subjek A: “Penawaran khusus hanya untuk Anda” menghasilkan tarif terbuka sebesar 5% dibandingkan dengan baris subjek B: “Jangan lewatkan kesempatan ini” sebesar 3%.

Apakah itu berarti baris subjek A lebih baik daripada baris subjek B? Itu tergantung pada margin kesalahan Anda. Jika margin kesalahan adalah 2%, maka tingkat terbuka aktual atau interval kepercayaan baris subjek A berada di antara 3% dan 7%. Karena ujung bawah interval tumpang tindih dengan hasil baris subjek B sebesar 3%, Anda tidak dapat menyimpulkan bahwa terdapat perbedaan yang signifikan secara statistik antara baris subjek A dan B. Memeriksa margin kesalahan penting saat membuat kesimpulan berdasarkan pengujian Anda hasil.

**Ingin menghitung margin kesalahan Anda?**

Yang Anda butuhkan hanyalah ukuran populasi, tingkat kepercayaan, dan ukuran sampel. Untuk lebih memahami kalkulator ini, tinjau istilah-istilah berikut:

* **Tingkat kepercayaan** : Persentase yang menunjukkan seberapa besar kemungkinan sampel Anda secara akurat mencerminkan populasi yang lebih besar
* **Populasi** : Jumlah total sampel yang Anda ambil
* **Sampel** : Bagian dari suatu populasi yang mewakili populasi tersebut
* **Margin of error** : Jumlah maksimum hasil sampel yang diharapkan berbeda dari populasi sebenarnya

Dalam kebanyakan kasus, tingkat kepercayaan 90% atau 95% digunakan. Namun, bergantung pada industri Anda, Anda mungkin ingin menetapkan tingkat kepercayaan yang lebih ketat. Tingkat kepercayaan 99% merupakan hal yang wajar di beberapa industri, seperti industri farmasi.

Setelah Anda menentukan ukuran populasi, ukuran sampel, dan tingkat kepercayaan, masukkan informasi tersebut ke dalam kalkulator margin kesalahan seperti di bawah ini:

* [Kalkulator margin kesalahan oleh Good Kalkulators (kalkulator online gratis)](https://goodcalculators.com/margin-of-error-calculator/)
* [Kalkulator margin kesalahan oleh CheckMarket](https://www.checkmarket.com/sample-size-calculator/#sample-size-margin-of-error-calculator)

**Pengambilan kunci**

Margin of error digunakan untuk menentukan seberapa dekat hasil sampel Anda dengan kemungkinan hasil jika Anda dapat mensurvei atau menguji seluruh populasi. Margin of error membantu Anda memahami dan menafsirkan hasil survei atau tes dalam kehidupan nyata. Menghitung margin kesalahan sangat membantu ketika Anda diberikan data untuk dianalisis. Setelah menggunakan kalkulator untuk menghitung margin kesalahan, Anda akan mengetahui seberapa besar perbedaan hasil sampel dari hasil seluruh populasi.