La importancia de la estadística y la replicación en microbiología

Abuela: ¡Juego mucho mejor si tomo una bebida energética en el entretiempo!

Foto de Kelly 1 de Pexels

Jaime I Prosser

Facultad de Ciencias Biológicas, Universidad de Aberdeen, Reino Unido:

Un marco educativo en microbiología centrado en la niñez

La importancia de la estadística y la replicación en microbiología

Sinopsis

Los estudios científicos tienen como objetivo explicar los fenómenos observados mediante hipótesis que generan predicciones que pueden comprobarse mediante la comparación con datos experimentales. Tanto la observación inicial de los fenómenos como la comprobación experimental implican la recopilación y el análisis de datos. Si bien estos datos pueden ser cualitativos, tanto la descripción de los fenómenos naturales como la comprobación crítica de las hipótesis son más minuciosas, precisas y completas si se emplean datos cuantitativos. Por ejemplo, podemos observar que la pasteurización reduce la abundancia bacteriana en la leche, pero es de mucho mayor valor si observamos que la pasteurización reduce la abundancia al 50%, al 1% o al 0,0001%. Por lo tanto, los datos cuantitativos son fundamentales para la microbiología y la estadística proporciona las técnicas para la organización, presentación, análisis e interpretación de estos datos. La estadística puede ser de dos tipos: *Estadística descriptiva* para resumir datos, visualmente o, por lo general, numéricamente. Por el contrario, *Estadística inferencial* Se utiliza para interpretar y extraer conclusiones de nuestros datos. Algunos estudios no requieren análisis estadístico, por ejemplo, si solo implican datos cualitativos. En la práctica, la mayoría genera datos cuantitativos y requieren análisis estadístico. Por lo tanto, es esencial que cualquier microbiólogo comprenda los principios subyacentes al análisis estadístico e identifique los métodos estadísticos necesarios para recopilar o analizar datos.

En el siguiente análisis se describen algunos de los principios básicos de la estadística aplicados a datos continuos en situaciones relativamente sencillas. No es posible analizar la amplia gama de datos microbiológicos encontrados, el diseño experimental complejo o todos los supuestos en los que se basan los análisis, pero los principios se aplican en todos los casos. El análisis estadístico implica necesariamente matemáticas y ecuaciones, que se presentan en recuadros. En el pasado, estos cálculos se realizaban con calculadoras u hojas de cálculo, pero ahora se utilizan habitualmente paquetes de software estadístico. Si bien esto es conveniente, a menudo necesario y genera información más detallada, presenta el peligro real de aceptar los resultados del software sin comprender los principios subyacentes y puede dar lugar a interpretaciones erróneas.

La microbiología y el contexto social

El análisis estadístico es necesario para cualquier estudio microbiológico que involucre datos cuantitativos. Por lo tanto, es transversal y necesario en todos los aspectos de la microbiología y los ODS.

1. Observaciones generales

a. *Poblaciones y muestras.* Al analizar datos, distinguimos entre una población, que es la colección más grande de entidades por las que tenemos interés, por ejemplo, todas las bacterias en un cultivo o varios cultivos cultivados en diferentes condiciones. Rara vez es factible medir las propiedades de cada miembro de una población y, por lo general, medimos las propiedades de una porción o muestra de la población, de la que inferimos las propiedades de la población. Al hacer esto, es importante muestrear las poblaciones de manera aleatoria para evitar sesgos, es decir, cada miembro de la población debe tener la misma probabilidad de ser elegido. La precisión y, en consecuencia, la cantidad de información que obtenemos, aumentan a medida que aumenta el número y el tamaño de las muestras.

b. *Variabilidad y error.* Se requiere un análisis estadístico debido a la variabilidad, por ejemplo. La abundancia bacteriana en la leche pasteurizada de diferentes tiendas o en diferentes envases

variará. Parte de esta variabilidad será *sistemática* (potencialmente explicables), mientras que algunos serán aleatorios o *error experimental*. Este último surge debido a la variabilidad inherente en el material experimental o la falta de uniformidad en la realización física del experimento. Ambos tipos de error deben minimizarse para mejorar la potencia de cualquier prueba estadística, por ejemplo, mediante el manejo del material experimental para reducir los efectos de la variabilidad inherente, refinando la técnica experimental o el sentido común.

c. *Replicación.* Una técnica para reducir el error es la replicación. Si un tratamiento aparece más de una vez en un experimento, se dice que se ha replicado. La replicación cumple dos funciones principales. En primer lugar, proporciona una estimación del error experimental, que es necesaria para las pruebas de significancia y límites de confianza. Si solo hay un tratamiento, es decir, una sola réplica, no hay información sobre el error experimental y es imposible determinar si las diferencias entre este tratamiento y otro se deben a diferencias en los tratamientos o a diferencias entre unidades experimentales. En segundo lugar, la replicación mejora la precisión de un experimento al reducir la desviación estándar de la media del tratamiento (ver a continuación).

El número de réplicas necesarias depende de la precisión requerida, que puede ser difícil decidir de antemano, pero no tiene sentido realizar un experimento que no proporcione la precisión requerida. Hay otras formas sofisticadas de reducir el error experimental, pero la mayoría son pura cuestión de sentido común. Es esencial eliminar las técnicas descuidadas, ya que a menudo no son aleatorias y están sesgadas y constituyen inexactitud en lugar de variabilidad.

d. *Precisión y exactitud.* La precisión y la exactitud se consideran sinónimos coloquialmente, pero tienen significados diferentes en el análisis estadístico. Ambos son medidas de error, pero la exactitud describe qué tan cerca están las observaciones de su valor "verdadero", mientras que la precisión describe qué tan cerca están las mediciones entre sí.

2. Estadísticas descriptivas

a. *Promedios.* Los promedios o medidas de tendencia central son formas importantes de describir los datos. Por ejemplo, si medimos la longitud de 100 celdas, podríamos presentar los datos como una lista o un histograma, pero es mucho más conveniente describirlos como un promedio. La medida más común del promedio es la media (Cuadro 1), que es fácil de calcular, estable a las fluctuaciones en el muestreo y susceptible de manipulación algebraica. Otro promedio es la mediana, el valor central cuando se ordenan todas las mediciones, que es más estable con respecto a los valores extremos.

b. *Variabilidad.* Una descripción más completa de los datos requiere una medida de dispersión o variabilidad, siendo las dos más comunes la varianza y la desviación típica. Podríamos cuantificar la variabilidad sumando la diferencia entre cada valor y la media, pero, por definición, esta suma sería = 0. Los cuadrados de cada diferencia serán positivos y la suma de estos cuadrados se denomina suma total de cuadrados corregida (*SS*). La SS aumenta con el tamaño de la población o de la muestra, por lo que debe estandarizarse para comparar poblaciones de diferentes tamaños. Para las poblaciones, esto se logra dividiendo *SS* por el tamaño de la población, *N*, pero para la media de la muestra dividimos por $n - 1$, donde n es el tamaño de la muestra (Cuadro 1). Esto reduce el sesgo y el exceso de confianza al estimar la variabilidad en la población a partir de la de una muestra. Este sesgo resulta del hecho de que ya hemos calculado la variabilidad total al calcular la media y, por lo tanto, podemos calcular la $n^{ésima}$ diferencia una vez que conocemos los otro $n - 1$ valores. Por lo tanto, decimos que este valor final "no es libre de moverse" y que tenemos $n - 1$ *grados de libertad.* Los valores que obtenemos se denominan varianza poblacional o muestral, pero sus unidades son el cuadrado de nuestras medidas, por ejemplo, cm^2, lo cual no es muy intuitivo, y es más habitual describir la variabilidad como la *desviación estándar*, que es la raíz cuadrada de la varianza (véase

el recuadro 1). La desviación estándar puede considerarse como la distancia promedio de los valores con respecto a la media.

c. *Coeficiente de correlación.* Las estadísticas anteriores son "uni variadas", es decir, implican una sola variable, por ejemplo, la altura. A menudo medimos dos o más variables en el mismo individuo, por ejemplo, la altura, el peso, la edad, que se describen mediante estadísticas bi variadas o multivariadas, respectivamente. Las asociaciones entre dos variables, "x" y "y", se pueden visualizar al representarlas gráficamente para cada individuo en los ejes "x" y "y". La fuerza de la asociación se puede cuantificar mediante el análisis de correlación y el cálculo de un coeficiente de correlación. Un ejemplo es la r de Pearson o correlación producto-momento (véase el Cuadro 1).

Las ecuaciones para los coeficientes de correlación son más complicadas (véase el recuadro 1), pero el numerador representa la covarianza, una medida de cómo varía x con y, calculada como la suma del producto de las diferencias entre cada variable y su media. Esto es equivalente a la varianza de los datos uni variados, mientras que el denominador la normaliza con respecto a las desviaciones estándar de cada variable.

El coeficiente de correlación cuantifica la *fortaleza* de la asociación entre dos variables y no tiene unidades, ya que representa la relación de varianzas. Tiene valores entre -1 (asociación negativa completa), pasando por 0 (sin asociación) hasta +1 (asociación positiva completa). Cuanto más cerca de -1 o +1, más fuerte es la asociación. No es una medida del cambio cuantitativo de "x" con respecto a "y". Lo más importante es que no proporciona información sobre causa y efecto. Por lo tanto, un coeficiente de correlación alto no significa que una variable esté afectando a la otra o que sea afectada por ella.

d. *Regresión lineal*. La regresión lineal cuantifica la relación entre una variable dependiente y y una variable independiente x. En esta situación, tenemos el control de x y asumimos que existe una relación lineal entre "y" y "x", por ejemplo, la relación entre la densidad óptica (y) y la concentración (x). La regresión lineal estima la fuerza de la relación y permite el cálculo de y para cualquier valor de x, pero solo dentro del rango de valores de x e y que haya medido. La relación entre x e y está representada por

$$y = a + bx + E$$

donde a y b son la intersección y la pendiente de la línea que minimiza el error (E), es decir, dan la línea de mejor ajuste. Las ecuaciones para la regresión lineal se dan en el Cuadro 1 y el software estadístico proporciona otras estadísticas sobre la calidad del ajuste, la confianza en a y b, etc.

3. ***Estadística inferencial***

a. Distribuciones. La media y la varianza son útiles para describir la "situación más probable". evento' y la variación en torno a él, pero son de mayor utilidad estadística si conocemos cómo se distribuyen los valores. Muchos caracteres biológicos continuos siguen la distribución gaussiana o normal (Fig. 1).

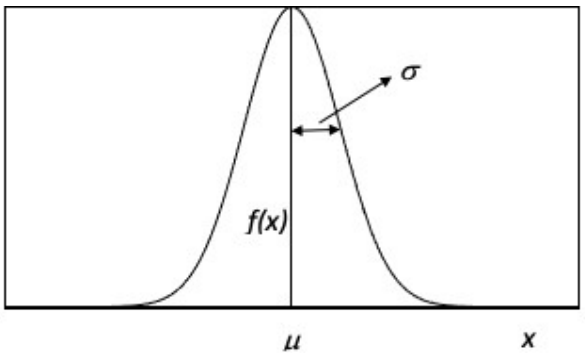

Fig. 1. Distribución normal de *f(x)*, la frecuencia de una variable *x*, en función de *x*.

Así que, si *x* representa la altura de las personas, y *f(x)* la frecuencia de ocurrencia de diferentes alturas, la gráfica de *f(x)*contra *x* dará una curva en forma de campana con media, μ, y desviación estándar, σ. Note que la variación es continua y se distribuye de manera uniforme alrededor de la media, es decir, la distribución es simétrica y las desviaciones de la misma son igualmente probables en cada dirección. En teoría, no hay límites superiores o inferiores para *x*, pero la frecuencia de los extremos es muy baja. La desviación típica mide la distancia desde la media hasta el punto de inflexión de la curva. (Debe recordarse, sin embargo, que no todas las cosas se distribuyen normalmente).

La distribución normal se puede utilizar para calcular la proporción de la población con características por encima o por debajo de un valor determinado. Sin embargo, estos cálculos no son sencillos sin el uso de computadoras y dependen de μ y σ. Antes de que el software informático estuviera fácilmente disponible, este problema se resolvía utilizando la distribución normal estándar o la distribución. Si restamos la media de la población de cada valor individual, obtendremos una media de 0. De manera similar, como la desviación estándar es la desviación promedio de la media, la división de cada diferencia por la desviación estándar dará como resultado una nueva desviación estándar de 1.

La estandarización se consigue entonces mediante la ecuación: $z = x - \mu / \sigma$. La distribución tiene la misma forma que la distribución normal, pero con una media de 0 y una desviación estándar de 1. En el pasado, esto permitía el cálculo de diferentes probabilidades (áreas bajo la curva) utilizando tablas de la distribución normal estándar, y esta transformación ahora hace que el cálculo sea más rápido. Tanto la distribución normal como -Las distribuciones se pueden describir mediante una ecuación (ver Cuadro 1) y aproximadamente el 68% de los valores caen dentro de la media ± desviación estándar, mientras que el 95% y el 99,7% caen dentro de 2 y 3 desviaciones estándar de la media (Fig. 2).

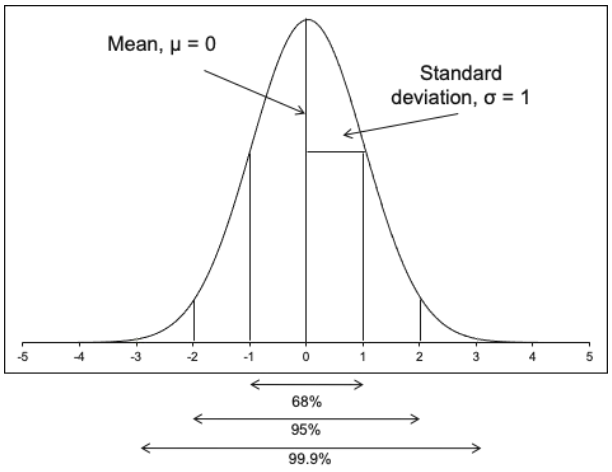

Fig. 2. Normal estandarizada o *z*-distribución, indicando las probabilidades (áreas bajo la curva) dentro de la media ± 1, 2 y 3 desviaciones estándar.

b. Significación estadística. La estadística inferencial nos permite sacar conclusiones de nuestros datos. Por ejemplo, podemos querer saber si las bacterias crecen de manera diferente en diferentes medios o a diferentes temperaturas. En efecto, esto es una prueba de hipótesis, en el sentido de que estamos probando la hipótesis de que el medio o la temperatura afectan el crecimiento. Cualquier prueba de una idea o hipótesis, ya sea a partir de un razonamiento teórico o sugerida por resultados de experimentos anteriores, debe implicar una declaración clara de objetivos y se define la hipótesis nula, a la que se le da el símbolo Ho. Estadísticamente, Ho es una declaración de "no hay diferencia" entre

dos poblaciones muestreadas que se puede esperar que difieran, por ejemplo, a través de diferentes tratamientos.

Las pruebas de significancia son la base del análisis estadístico inferencial y se discutirán inicialmente imaginando la situación en la que hemos muestreado dos poblaciones diferentes, *X* y *Y*, que sospechamos que son diferentes. Por lo tanto, estamos probando la hipótesis nula, Ho, que las medias poblacionales indican que son iguales, $\mu 1 = \mu 2$, pero podemos realizar la prueba poniendo énfasis en una hipótesis alternativa, HA, $\mu 1 > \mu 2$, ignorando cualquier evidencia de que $\mu 1 < \mu 2$. Tomamos muestras de cada población y calculamos las medias muestrales, $\bar{x}1$ y $\bar{x}2$. Si $\bar{x}1 > \bar{x}2$, entonces podríamos rechazar Ho en favor de HA. Sin embargo, este resultado puede deberse a que $\mu 1$ realmente es mayor que $\mu 2$o porque resulta que las muestras tomadas lo hicieron parecer así, debido a las fuentes no controladas de error y variación. Esto siempre es un riesgo al tomar muestras. Cuanto más pequeña sea la muestra y más pequeña la diferencia entre $\mu 1$ y $\mu 2$, mayor es el riesgo.

Para distinguir entre estas razones y elegir Ho o HA. Debemos introducir el concepto de significancia. Para ello calculamos la probabilidad de que una diferencia al menos tan extrema como $\bar{x}1 - \bar{x}2$ (la diferencia observada) surgió por casualidad bajo el supuesto de que Ho es cierto (es decir $\mu 1 = \mu 2$. Si la probabilidad es pequeña, podemos concluir con una certeza razonable (pero no absoluta) que la diferencia es real. Si la probabilidad es grande, concluimos que la diferencia no es real, sino que surgió del azar como resultado del muestreo y la experimentación.

c. Error estándar e intervalos de confianza. Ahora debemos considerar nuestra confianza en que la media de la muestra representa con precisión la media de la población. Podríamos tomar muchas muestras, cada una de las cuales daría una media diferente, y se puede demostrar que las medias de todas estas muestras se distribuyen normalmente y que la desviación estándar de la media, también llamada error estándar, viene dada por la ecuación $S_{\bar{x}} = s / \sqrt{n}$. Esta ecuación define, cuantitativamente, cómo la replicación aumenta nuestra confianza en que la media de la muestra refleja la media de la población, con la varianza disminuyendo la precisión aumentando en proporción a la raíz cuadrada del tamaño de la muestra.

Los datos a menudo se presentan como la media ± desviación o error estándar. Una alternativa es utilizar límites de confianza que contendrán un parámetro con una probabilidad del 95 % (o algún otro valor). Estos se denominan límites de confianza del 95 % y, para muestras de gran tamaño, se pueden determinar utilizando la distribución normal estandarizada, en función de la media y la desviación estándar de nuestra muestra: $z = (x - \bar{x}) / s$. Ahora, queremos saber el valor de *X*. Esto dará una probabilidad del 2,5 %. Este valor es 1,96 (ver figura 2), por lo que la ecuación se convierte en: $X = \bar{x} \pm 1{,}96s$. Si el tamaño de la muestra es relativamente pequeño (<40), utilizaríamos el valor equivalente de *a*- distribución (ver más abajo) en lugar de la distribución normal.

d. *Prueba t del estudiante.* Supongamos que queremos probar una teoría que predice que el diámetro de una hifa de hongo, en determinadas condiciones, tendrá un tamaño de 11 µm, es decir, que la media poblacional, µ = 11 µm. En términos estadísticos, estamos probando la hipótesis nula Ho: µ = µo= 11 y medimos el diámetro de 100 hifas y calculamos una media muestral de 11,14 y un error estándar de 0,11 µm. Luego transformamos nuestra variable a la distribución normal estandarizada, es decir, si Ho es cierto, entonces $z = (\bar{x} \sim 11) / 0{,}11$ se distribuye normalmente con media 0 y varianza 1.

Evaluamos la importancia del valor observado $\bar{x} = 10{,}86$ calculando la probabilidad de que un valor medio al menos tan extremo como 10,86 pueda ocurrir por casualidad, suponiendo Ho es cierto. Cualquier valor menor a 10,86 es ciertamente más extremo que la *X* observada, pero también lo es cualquier valor mayor que 11,14, es decir, debemos considerar los extremos por encima o por debajo de la media y determinar la probabilidad de que $\bar{x} < 10{,}86$ y $\bar{x} > 11.14$. Esta probabilidad (calculada a partir de tablas de distribución z o de un programa informático) es 2 x 0,102 = 0,204. Esto significa que si Ho. Es cierto que esperaríamos una media muestral de 10,86 de aproximadamente el 20%, una

muestra de cada cinco. Por lo tanto, no podemos concluir que Ho es falso, es decir μ = 11 con una probabilidad de 1 en 5. Por convención, empezamos a dudar de Ho. Cuando la probabilidad alcanza aproximadamente 1 en 20, o 5% (0,05). Esto a veces se describe como el nivel de significación del 5% o 0,05.

Ninguna hipótesis nula puede considerarse aisladamente. Siempre existe una hipótesis alternativa, HA, aunque no se indique explícitamente. Más arriba probamos Ho: $\mu = \mu_o$ contra HA: $\mu \neq \mu o$ es decir $\mu < \mu o$. Por lo tanto, cuando consideramos los valores de $\bar{x}$ al menos tan extremo como la observada $\bar{x}$ = 10,86 esto incluye cualquier valor < 10,86 pero también cualquier valor > 11,14. Esto se llama prueba de dos colas. Si hubiéramos declarado HA: $\mu < \mu o$ entonces cualquier $\bar{x}$ > 11 no habría sido extrema, ya que habría sido más probable que viniera de una población con μ = 11 que con μ < 11. Entonces, en este caso utilizaríamos una prueba de una cola y habríamos obtenido una probabilidad de 0,102.

Otra consideración es el tamaño de la muestra. Con tamaños de muestra grandes, $\bar{n}$ > 30-40, $\bar{x}$ y *s* son estimaciones razonablemente precisas de μ y σ. Sin embargo, para tamaños de muestra más pequeños, no podemos confiar en la precisión de s^2, en lugar de utilizar la distribución z, utilizamos la distribución t de Student. Realizamos una transformación similar: $t = (\bar{x} - \mu_o) / s_{\bar{x}}$ y asumir que *t* sigue l*a* *t*-distribución con *n* - 1 grado.

Obsérvese que, en ambos ejemplos, la estadística que se calcula contiene la "diferencia" que se investiga, mientras que el denominador es una medida del error experimental. En otras palabras, estamos determinando la relación entre una diferencia observada debida a un efecto potencial y el error experimental. Este enfoque se aplica en la situación posiblemente más común de comparar dos medias de muestra, en lugar de comparar una media con un valor fijo. En este caso, el numerador es la diferencia entre las medias de muestra, mientras que el denominador es una medida del error experimental combinado. Para un tamaño de muestra pequeño y diferentes tamaños de muestra, la estimación de la ecuación de la varianza común se vuelve más complicada (Cuadro 1).

e. *Análisis de varianza.* En muchos experimentos queremos comparar varias medias, en lugar de solo dos, por ejemplo, podríamos estar investigando el efecto de 5 medios de cultivo diferentes en el rendimiento de biomasa, con una hipótesis nula de que el medio no tiene efecto en la biomasa. Tenemos 4 réplicas para cada medio, lo que da un total de 20 mediciones de biomasa, y calculamos 5 valores medios, cada uno a partir de 4 réplicas. Para evaluar si el medio de cultivo afecta el rendimiento, podemos realizar un análisis de varianza.

El primer paso es evaluar la media y la varianza en las 20 culturas, esta última siendo la suma total corregida de los cuadrados (ver arriba). El análisis de varianza divide esta varianza total en una serie de componentes que creemos que están relacionados con diferentes circunstancias causales (tratamientos o factores), en nuestro caso el uso de diferentes medios y el error experimental. Calcula las varianzas sobre las medias de estos componentes y evalúa la significancia de estas varianzas.

En nuestro caso, tenemos dos fuentes de variación: el tratamiento (medio de cultivo) y la variación inherente (error experimental). El error experimental se puede estimar calculando, de forma independiente, la suma de cuadrados dentro de cada tratamiento y luego sumando estos valores para obtener la "suma de cuadrados dentro de cada tratamiento", SSwit. Si el medio de crecimiento afecta la biomasa, entonces *SSwit* será menor que *SStot*, y la varianza restante debida al tratamiento, es decir, la varianza entre tratamientos o *SSbet*.

Siempre habrá alguna variación aleatoria, pero si esta es mayor de lo que esperaríamos Entonces, puede que no haya una diferencia real entre los tratamientos. Por lo tanto, elaboramos una tabla de análisis de varianza (AOV) (Cuadro 2) que contiene la suma de cuadrados. También contiene los cuadrados medios intra e inter, que son la suma de cuadrados dividida por el número respectivo de grados de libertad. Si Ho es cierto y no hay efecto del tratamiento, esperaríamos que *EM* Los valores

deben ser aproximadamente iguales, ya que ambos estimarán la varianza total. Por lo tanto, calculamos la relación de los valores de MS (la relación de varianza) y la comparamos con los valores tabulados de la *F*- distribución para una probabilidad de 0,05 con los grados de libertad apropiados, o utilizar un software estadístico para generar un valor. Los valores de la distribución F se basan en grados de libertad asociados tanto con la variación entre variables como con la variación interna.

Participación de los alumnos

1. ***Discusión en clase sobre la importancia de la replicación y el análisis estadístico que sustenta descubrimientos científicos reportados en las noticias.***

2. ***Ejercicios***

a. Se pueden calcular estadísticas descriptivas básicas para las características de los alumnos, por ejemplo, altura, peso. El efecto del tamaño de la muestra y la distribución normal estandarizada se pueden ilustrar midiendo la altura en muestras de diferente tamaño y se pueden calcular los coeficientes de correlación para la altura y la edad. Las estadísticas inferenciales se pueden ilustrar mediante la comparación de las medias de diferentes grupos dentro de una clase o entre diferentes clases.

b. La pandemia de covid ha puesto de relieve la necesidad de realizar análisis microbiológicos cuantitativos. Datos y limitaciones en la comprensión básica de estos datos y los análisis estadísticos empleados. Por lo tanto, los datos estadísticos publicados sobre la COVID se pueden utilizar para ilustrar los principios básicos de las estadísticas y las formas en que los datos se pueden interpretar y malinterpretar.

Lectura adicional

Hay muchos libros introductorios sobre estadística, incluidos aquellos para quienes estudian biología y microbiología, y sitios web igualmente buenos, un buen ejemplo es https://www.scribbr.com/category/statistics/ .

Cuadro 1. Ecuaciones estadísticas a las que se hace referencia en el texto.

Media poblacional $\mu=\frac{\sum X}{N}$, donde N es el tamaño de la población y X es la medida individual Media muestral $\bar{x}=\frac{\sum x}{n}$, donde n es el tamaño de la población y x es la medida individual
Variabilidad: Población: Suma total corregida de cuadrados $¿\sum (X-\mu)^2$, varianza $\sigma^2=\frac{\sum (X-\mu)^2}{N}$, desviación estándar $\sigma=\sqrt{\frac{\sum (X-\mu)^2}{N}}$ Muestra: Suma total corregida de cuadrados $¿\sum (x-\bar{x})^2$, varianza $s^2=\frac{\sum (x-\bar{x})^2}{N}$, desviación estándar $s=\sqrt{\frac{\sum (x-\bar{x})^2}{n}}$
Coeficiente de correlación $r_{xy}=\frac{cov(x,y)}{s_x s_y}$ = $r_{xy}=\frac{n\sum XY-\sum X\sum Y}{\sqrt{¿¿¿}}$, donde X y Y son medidas individuales de dos variables, X y Y.
Regresión lineal Pendiente $b=\frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2}$, intercepta a = ȳ - b$\bar{x}$, dónde x y y son medidas individuales de una variable independiente x y una variable dependiente y.

Prueba t para la comparación de medias con tamaños de muestra desiguales

t-estadística $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n}+\frac{1}{m}\right)}}$, donde varianza común $s_p^2 = \frac{\sum x_1^2 - \frac{\left(\sum x_1\right)^2}{n} + \sum x_2^2 - \frac{\left(\sum x_2\right)^2}{m}}{n+m-2}$ y $\bar{x}_1 \wedge \bar{x}_2$ son medias de dos muestras de tamaño n y m.

Cuadro 2. Tabla de análisis de varianza.

Fuente de variación	SS corregido	*df*	*MS*	Coeficiente de varianza
Entre (tratamiento)	SS_{bet}	k - 1	$\frac{SS_{bet}}{k-1}$	$\frac{MS_{bet}}{MS_{wit}} = F_{(k-1, N-1)}$
Dentro de (error)	SS_{wit}	$(N-1) - (k-1)$	$\frac{SS_{wit}}{(N-1)-(k-1)¿}¿$	
Total	SS_{tot}	N - 1		