Estadística (Q)
Aplicaciones de algunos temas de la materia a la química.

1   Variables Poisson y exponenciales

La siguiente es una colaboración del Dr. Pedro Aramendía. Seleccionamos dos ejercicios de la materia Cinética Química que ejemplifican bien la relación entre las distribuciones Poisson y exponencial. Notar que el lenguaje de los químicos muchas veces no coincide con el de los estadísticos (aunque uds. ya deben saberlo). Parte importante del ejercicio es aprender a fusionar los dos lenguajes.

1. Un sistema de iluminacin irradia sobre un blanco 10^9 fotones por segundo. Genere una posible secuencia de eventos (fotones emitidos) en 1 μs.

   (a) Grafique en funcin del tiempo el número de fotones irradiado en intervalos consecutivos de 5 ns.

   (b) Grafique la distribución del nmero de fotones por intervalo y compruebe que es Poissoniana.

   (c) Grafique la distribución de tiempos entre eventos y compruebe que es exponencial.

   (d) Calcule el número de fotones que deber irradiarse en un segundo para que la desviación estándar relativa porcentual (desviación estándar / promedio 100) del número de fotones incidentes en 5 ns sea menor que el 1 %.

   (e) Calcule el nmero de fotones que deberá irradiarse en un segundo para que la probabilidad de apilamiento en 50 ps sea menor que el 1 %.

   Los siguientes gráficos representan una posible secuencia y las distribuciones correspondientes.
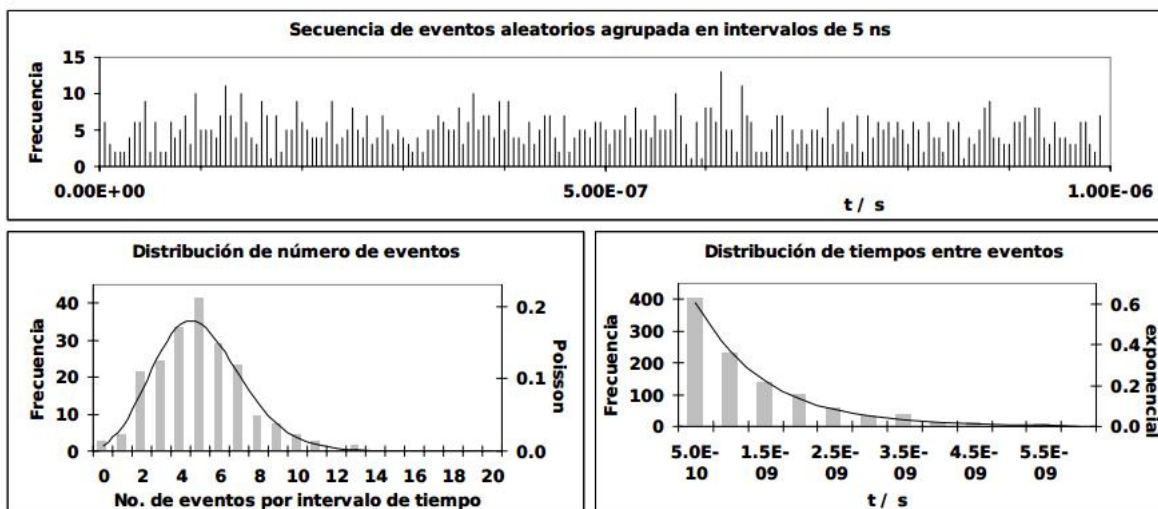
2. El gráfico representa la probabilidad de detección de un débil pulso de luz por diferentes seres humanos en función del nmero de fotones que llegan a la retina (W. Bialek, `http://www.princeton.edu/~wbialek/PHY562/1.pdf`; S. Hecht, S. Shlaer, M.H. Pirenne, J. Gen. Physiol. 25, 1942, 819-840).
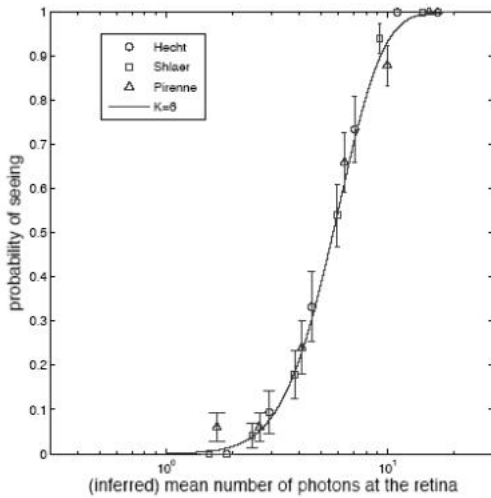
(a) Considerando que el número de fotones que inciden en la retina responde a la distribución de Poisson, y que la retina debe recibir un mínimo de 6 fotones para percibir el pulso, encuentre la relación matemática existente entre las variables del gráfico.

(b) La visión del ojo humano en la oscuridad se debe a la alta sensibilidad de los bastones contenidos en la retina. Teniendo en cuenta que en la percepción del pulso de luz en las condiciones de la experiencia participan alrededor de 500 bastones, calcule la relación entre las probabilidades de que en un mismo bastón se absorban uno o dos fotones cuando el nmero total de fotones absorbidos es 6.

(c) De acuerdo con los resultados anteriores evalúe la validez del siguiente enunciado: Para que un pulso de luz sea visible es necesario estimular al menos seis bastones; para ello es suficiente que cada uno absorba un solo fotón.

2 Intervalos de confianza

1. Most of the time the population mean differs from the mean of each individual sample taken from the population. Consider the following example where absorbance of a solution containing a known concentration of substance A was determined by U.V./Visible spectrometer. The absorbance of the solution was measured three times during each experiment and the average value, standard deviation s and 2s was calculated (see table below). The mean values from Figure 1 were plotted in the graph shown in

Measurement	EXP1	EXP2	EXP3	EXP4	EXP5	EXP6	EXP7	EXP8	EXP9	EXP10	EXP11	EXP12	EXP13
1	0.6006	0.6002	0.601	0.6006	0.5991	0.5998	0.6	0.6	0.6012	0.603	0.5997	0.601	0.598
2	0.5991	0.5993	0.5995	0.6002	0.5993	0.5995	0.6001	0.5997	0.6011	0.6035	0.5995	0.6005	0.5976
3	0.5998	0.5995	0.6001	0.601	0.5995	0.6001	0.6002	0.5999	0.6013	0.6025	0.5999	0.6015	0.5984
Mean	0.5998	0.5997	0.6002	0.6006	0.5993	0.5998	0.6001	0.5999	0.6012	0.6030	0.5997	0.6010	0.5980
s	7.51E-04	4.73E-04	7.55E-04	4.00E-04	2.00E-04	3.00E-04	1.00E-04	1.53E-04	1.00E-04	5.00E-04	2.00E-04	5.00E-04	4.00E-04
2s	1.50E-03	9.45E-04	1.51E-03	8.00E-04	4.00E-04	6.00E-04	2.00E-04	3.06E-04	2.00E-04	1.00E-03	4.00E-04	1.00E-03	8.00E-04

Figure 1: Absorbance values measured for a solution A using a U.V./Visible spectrometer. Three consecutive measurements were recorded during each experiment.

Figure 2 below and 2s (2 * standard deviation s) is shown for each mean. As can you see, the mean value calculated for the absorbance of A in each experiment (sample of the population) differs from the mean of the population (red line). Please also note that the interval - solid line above and below the mean in each experiment (Fig. 2) - of each mean (the range of values the mean can take with a certain probability) does not always contain the population mean (experiments 5, 9, 10, 13).
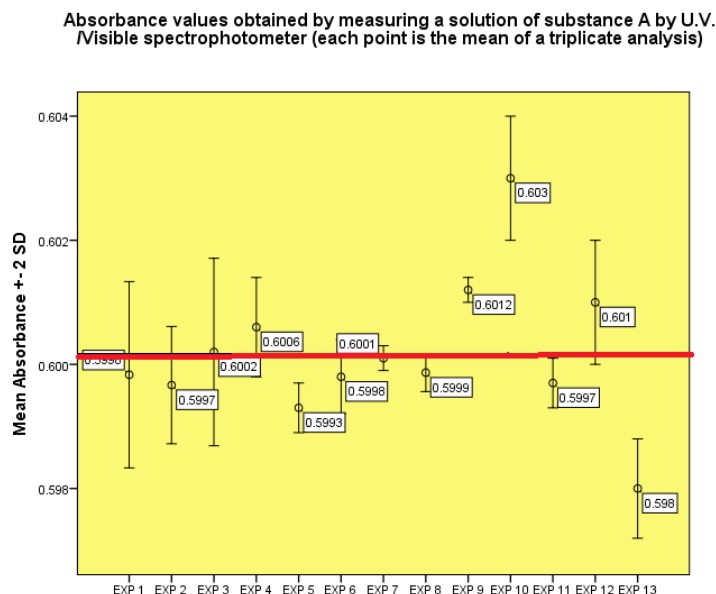

Figure 2: Absorbance values obtained by measuring a solution of substance A with known concentration C using a U.V/Visible spectrometer. Each point shown on the graph is the mean of three consecutive measurements of the solution of substance A with concentration C.

3 Tests de hipótesis

1. **Una forma gráfica de explicar los errores de tipo I y II (del blog de Eric Cai)**

   In hypothesis testing,

   - A Type 1 error is the rejection of the null hypothesis when it is actually true
   - A Type 2 error is the acceptance of the null hypothesis when it is actually false. (Some statisticians prefer to say failure to reject rather than accept the null hypothesis for Type 2 errors.)

   A Type 1 error is also known as a false positive, and a Type 2 error is also known as a false negative. This nomenclature comes from the conventional connotation of

   - The null hypothesis as the negative or the boring result
   - The alternative hypothesis as the positive or exciting result.

   A great way to illustrate the meaning and the intuition of Type 1 errors and Type 2 errors is the following cartoon.
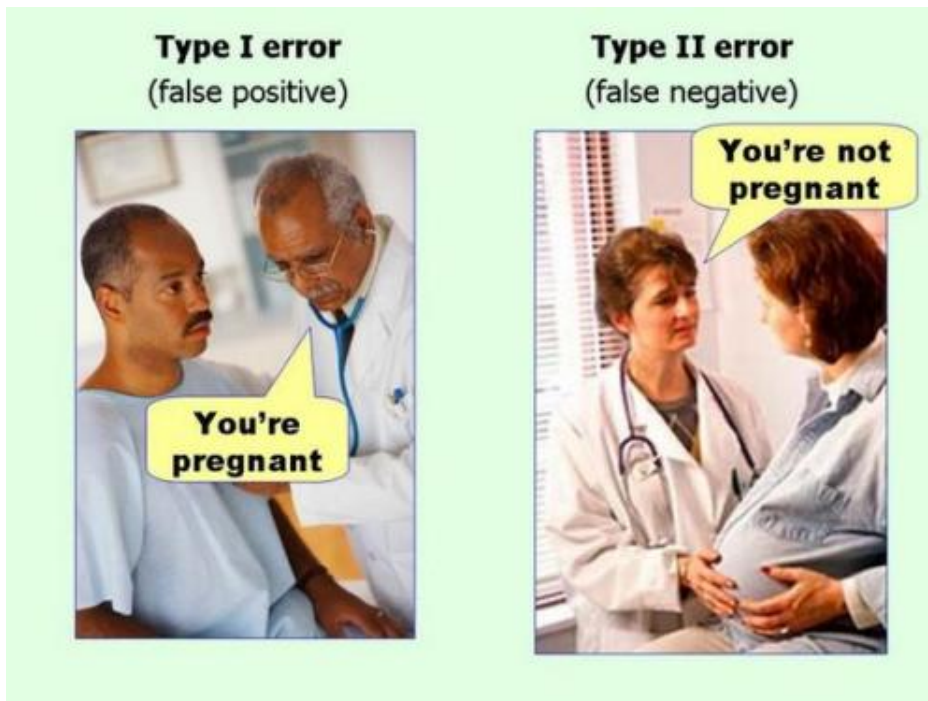
2. Juanito y el lobo

Statisticians have a different version of
The Boy Who Cried Wolf.

3. **Descubriendo el Argon con el Test de Welch para 2 muestras (del blog de Eric Cai).**

I learned about Lord Rayleighs discovery of argon in my 2nd-year analytical chemistry class while reading Quantitative Chemical Analysis by Daniel Harris. (William Ramsay was also responsible for this discovery.) This is one of my favourite stories in chemistry; it illustrates how diligence in measurement can lead to an elegant and surprising discovery. I find no evidence that Rayleigh and Ramsay used statistics to confirm their findings; their paper was published 13 years before Gosset published about the t-test. Thus, I will use a 2-sample t-test in R to confirm their result.
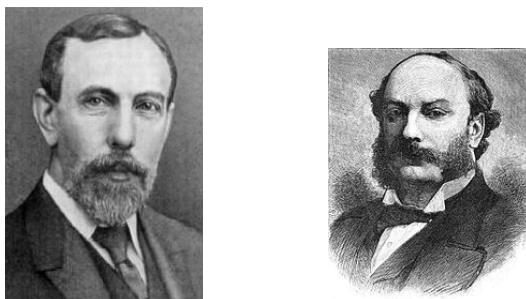

Figure 3: William Ramsay a la izquierda, Lord Rayleigh a la derecha

Rayleigh and Ramsay

John Williams Strutt (1842-1919), better known as Lord Rayleigh, was a very accomplished physicist and chemist known for Rayleigh scattering (which explains why the sky is blue), Rayleigh waves (surface acoustic waves in elastic solids) and the Rayleigh-Jeans law (which attempted to describe blackbody radiation but failed at short wavelengths, resulting in the ultraviolet catastrophe and motivating the development of quantum mechanics). Out of all of his accomplishments, he was recognized for the discovery of argon with the Nobel Prize in Physics in 1904.

William Ramsay (18521916) was a chemist at University College London, and his research focus was on nitrogen oxides before he learned about Rayleighs work and shared his curiosity about the discrepancy in the two masses. He and Rayleigh collaborated significantly on the discovery of argon, and he later won the Nobel Prize in Chemistry in 1904 for discovering the noble gases.

I could only find documentation about Rayleighs experimentation and data on the discovery of argon, but it is clear that Ramsay was responsible for this discovery, too, and the two scientists communicated many times about their work on this same problem. Even though I will refer to Rayleighs work and data in this blog post, Ramsay should also be recognized.

The Nitrogenous Origin

The discovery of argon can be traced back to Rayleighs original intent of measuring the densities of various gases in 1882. His first publication on this work appeared in 1888, and it discussed the relative densities of oxygen and hydrogen. He later focused on the density of gaseous nitrogen, which he measured in 2 ways.

1) Scientists knew that air contained roughly 75% nitrogen and 25% oxygen, so Rayleigh sought to obtain nitrogen by passing air over hot copper, removing oxygen via the reaction

$$O_2(g) + 2Cu(s) \rightarrow 2CuO(s)$$
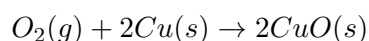
2) Rayleigh also produced nitrogen by bubbling air through liquid ammonia and then through a hot tube.

$$3O_2(g) + 4NH_3(l) \rightarrow 6H_2O(l) + 2N_2(s)$$
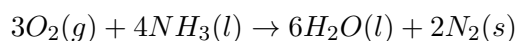
The water was then removed with a drying agent.

The Unexpected Result

Rayleigh discovered that the nitrogen from Method 1 (atmospheric nitrogen) was 2.3 mg heavier than the nitrogen from Method 2 (chemical nitrogen). While that may seem very small and quite possibly due to random error, he was confident that there was something in atmospheric nitrogen that caused this difference. Rayleigh sent a letter to Nature describing these results, and William Ramsay responded to Rayleigh with similar puzzlement.

Thus, Rayleigh decided to make nitrogen from the two sources with slightly different methods.

1) In addition to passing air over copper, Rayleigh also isolated nitrogen by

- passing air over hot iron to remove the oxygen
- passing air over freshly precipitated ferrous hydrate, or iron (II) hydrate, to remove oxygen

2) In addition to passing air over ammonia, Rayleigh also used chemical decomposition of the following nitrogenous compounds to isolate nitrogen:

- nitric oxide
- nitrous oxide
- urea
- ammonium nitrite

Exploring the Data

Here are the mass data from the two types of nitrogen, atmospheric and chemical, as entered into and displayed in R. Note that these are not paired data; 7 data were collected for nitrogen from air, and 8 data were collected for nitrogen from chemical decomposition. I created a data table called nitrogen.masses; note the use of the colnames() function to set the column names.

```
##### Analyzing Lord Rayleigh's Data on Nitrogen and Discovering Argon with the 2-Sample t-Test
##### Written by Eric Cai - The Chemical Statistician

atmospheric.nitrogen = c(2.31017, 2.30986, 2.31010, 2.31001, 2.31024, 2.31010, 2.31028, NA)
chemical.nitrogen = c(2.30143, 2.29890, 2.29816, 2.30182, 2.29869, 2.29940, 2.29849, 2.29869)
nitrogen.masses = data.frame(atmospheric.nitrogen, chemical.nitrogen)
colnames(nitrogen.masses) = c('Nitrogen from Air', 'Nitrogen from Chemical Decomposition')

> nitrogen.masses
        Nitrogen from Air       Nitrogen from Chemical Decomposition
1       2.31017                 2.30143
2       2.30986                 2.29890
3       2.31010                 2.29816
4       2.31001                 2.30182
5       2.31024                 2.29869
6       2.31010                 2.29940
7       2.31028                 2.29849
8       NA                      2.29869
```

It is easy to see that the masses in the first column are all slightly larger than the masses of the second column but by how much? A box plot is a useful way to visualize this contrast; it shows the following summary statistics of a set of data:

- Max{Min, Q1 1.5IQR}, where Min is the minimum, Q1 is the 1st quartile, and IQR is the interquartile range
- 1st quartile (25th percentile)
- median (50th percentile)
- 3rd quartile (75th percentile)
- Min{Max, Q3 + 1.5IQR}, where Max is the maximum, Q3 is the 3rd quartile, and IQR is the interquartile range

Here is the box plot.
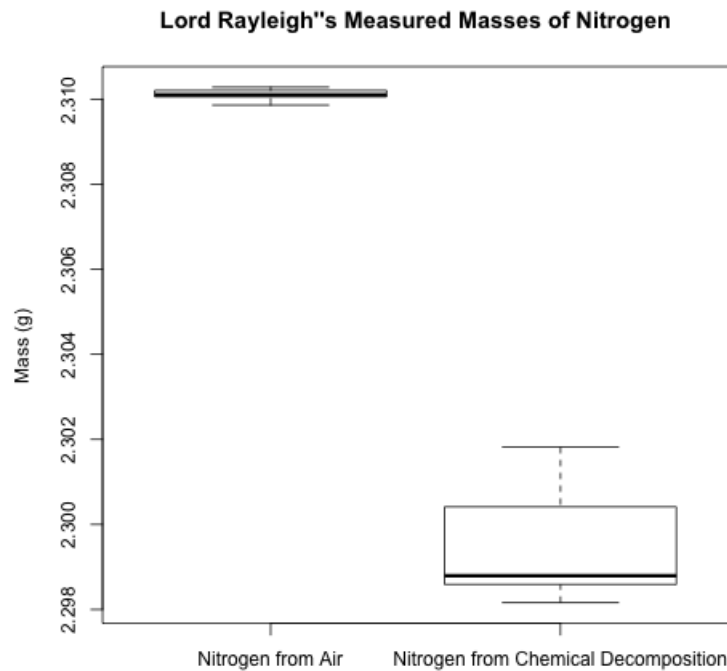
The 2-Sample t-Test to Compare Population Means

The box plot makes the difference between the masses of the 2 types of nitrogen even more striking. However, lets use a hypothesis test to compare the 2 population means to quantify just how certain we are that the difference is beyond just random error. Since we are comparing two population means, a 2-sample t-test is appropriate.

William Gosset (a.k.a. Student) developed the t-test in 1908, while Rayleigh and Ramsay published their discovery of argon in 1895, so Rayleigh and Ramsay could not have used the t-test for their data. In reading through Rayleighs and Ramsays paper, I actually dont find any statistical analysis to ensure that the difference was significant. Now that we do know about the t-test, lets use it to confirm their finding.

We are testing the following hypotheses, which are mutually exclusive:

Null Hypothesis: There is no difference between the masses of the 2 types of nitrogen.

Alternative Hypothesis: There is a difference between the masses of the 2 types of nitrogen.

Here is my R code for conducting the t-test, using the t.test() function.

```
> t.test(atmospheric.nitrogen, chemical.nitrogen)
```

8

```
Welch Two Sample t-test

data:  atmospheric.nitrogen and chemical.nitrogen
t = 21.5183, df = 7.168, p-value = 8.919e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
0.009495066    0.011827076
sample estimates:
mean of x    mean of y
2.310109      2.299448
```

The key result in the above output is the p-value, which is the strength of the evidence for the null hypothesis. A lower p-value indicates less evidence for the null hypothesis or, equivalently, more evidence for the alternative hypothesis. In this case, the p-value is 8.919 e-08, which is extremely low. (The usual cut-off is 0.05, which results in 95% confidence. Our low p-value gives much greater confidence!) Thus, the null hypothesis is rejected, and we have shown from Rayleighs data that the two samples of nitrogen are different!

Of course, this doesnt prove that the higher mass in atmospheric nitrogen is caused by the presence of an inert, gaseous element. (Yes, I realize that the title of my blog post is slightly misleading. :)) A series of attempts to characterize the inert residue from the atmospheric residue via chemical reactions yielded no results, and the emission spectrum of this residue was unlike any other known to chemists at the time. Rayleigh and Ramsay knew that claiming this inert gas as a new element was going to be controversial, since there was no place on the periodic table for it yet. Eventually, Ramsay discovered many other gases that were also chemically inert, and the noble gases were finally discovered as a new family on the periodic table.

References

- Daniel C. Harris. Quantitative Chemical Analysis, 7th Edition, Page 60-62.

- Russell D. Larsen. Lessons learned from Lord Rayleigh on the importance of data analysis. Journal of Chemical Education. 67, no. 11 (1990): 925.

- Carmen Giunta. Using history to teach scientific method: The case of argon. Journal of Chemical Education. 75, no. 10 (1998): 1322.

- Rayleigh, Lord, and William Ramsay. Argon, a New Constituent of the Atmosphere. Proceedings of the Royal Society of London 57.340-346 (1894): 265-287.

4 ANOVA

1. **Un artículo de Roland F. Hirsch (1977) sobre aplicaciones de ANOVA en química analítica y varias extensiones.**

   `http://pubs.acs.org/doi/pdf/10.1021/ac50016a715`

2. **Un artículo (dividido en 2) de Eric Cai sobre una aplicacion de ANOVA en papas fritas.**
   (notar que lo implementan en JMP, una plataforma quizás un poco más amigable que R).

   `https://community.jmp.com/t5/JMP-Blog/Potato-chips-and-ANOVA-in-analytical-chemistry-Part-1-For`
   `ba-p/30656`
   `https://community.jmp.com/t5/JMP-Blog/Potato-chips-and-ANOVA-Part-2-Using-analysis-of-variance-`
   `ba-p/30660`