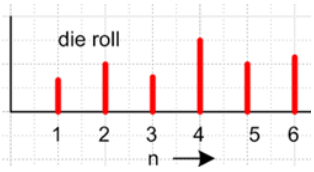
Appendix A Statistics PhL 2.8.15

Had I done more in the AFIB docs with error bars, I would have added this as an appendix, but I decided not to alter the has bled curve, so this will just sit here.

For any study one needs to have some idea of how believable the results of the study might be. In most study reports, each result is stated with two other numbers which define a 95% confidence interval for that result. In this appendix we discuss how that interval might be determined. In order to do this, we have to first provide a basic review of the machinery involved.

The subject hinges on the notion of an "experiment" which has "outcomes", but we shall confine our interest to a single outcome x which is a real number. We repeat the experiment N times, each time recording the outcome x. We assume that none of these experiments influences the others. We then plot the outcomes on a histogram and the shape of that graph gives us some idea of the underlying probability function p(x) which describes the outcome x. From this histogram we can compute things like the mean value of x which is usually written $\mu = \Sigma_x\, p(x)x$ and the variance $var = \Sigma_x\, p(x)(x-\mu)^2$. The variance is a measure of how spread out the distribution is about the mean. One often uses $\sigma \equiv \sqrt{var}$ since it has the same dimensions as x, and this σ is called the standard deviation of the distribution p(x). In general we don't know what p(x) is, but from doing N experiments and creating that histogram for x, we can get some idea of p(x), μ and σ. We suspect that by making N a large number, our estimates of p(x), μ and σ will be much better than for a small N.

Die Example. If the experiment is the roll of one die, and if the outcome x is the value of the number on the up-facing die surface, and if this experiment is performed N times (or if N dice are thrown one time), one might get a histogram like that shown below,

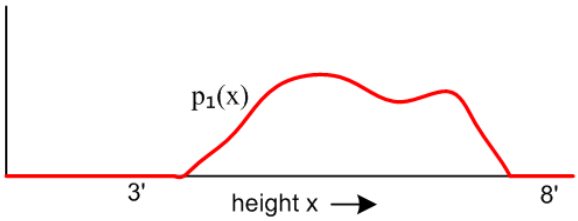


If N is very large, we would conclude from this histogram that the die was "weighted" and was not a "fair die". For a fair die, we expect all 6 bars to have the same height. Notice that the p(x) implied by the above chart is not symmetric, the mean is perhaps $\mu = 3.8$, and σ is some number in the range if 2. If the bars *were* the same height, one would have $\mu = 3.5$, $var = 2.92$ and $\sigma = \sqrt{2.92} = 1.7$ (just use the formulas shown above with $p(x) = 1/6$). On each throw of a die, the outcome x varies, so x is a variable (not a constant). Even though the die is weighted in the above example, x is still called a "random variable", even though one might feel that that weighted die has a non-random behavior. This word "random" just means the outcome x has some probability curve p(x) which has some spread to it ($\sigma > 0$).

If one writes x = X(s), one can make the distinction between X being a function and x being a particular value which the function takes. As s varies, X is always the same function, but the values x vary. It is in this sense that one often sees a random variable indicated by a capital letter, and its values in

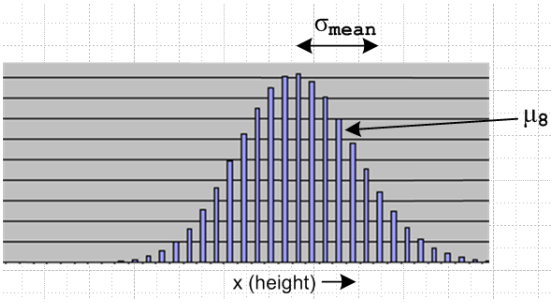
lower case. The official "random variable" is X, its values are x. For our situation where the experimental outcome x was assumed to be a real number, this distinction has no real significance.

Height Example. Now the experiment is "a person", and the outcome is x = the measured height of that person. The experiment is repeated with N people (not the same person N times). Now instead of assuming that N is "very large", we assume it is some moderate value. The result of our experiment is some continuous histogram (since x = height is a continuous variable),



Since this experiment was done with some moderate number of people N, $p_1(x)$ is only an approximation to the true underlying p(x) which we don't known. This curve has some mean μ_1 and some standard deviation σ_1 which we compute from $\mu_1 = \int dx\ p_1(x)x$ and $\sigma_1^2 = \int dx\ p_1(x)\ (x-\mu_1)^2$.

We would like to know just how close μ_1 might be to the true μ = average height of the population of all humans. In other words, how good was our height study? As a thought experiment only, suppose we were to repeat this trial of N people M different times (different people each time), and for each of these M trials we get a mean μ_i where i = 1,2..M. We could then plot all the different μ_i values on their own little histogram,

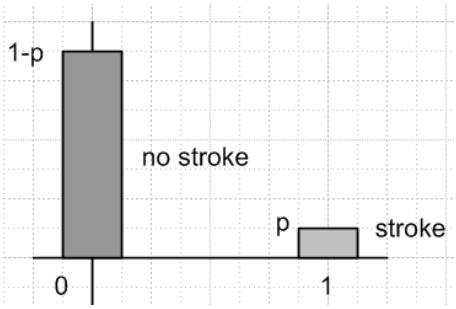


For illustration purposes, we drew this as a normal distribution shape, but that is not important. We presume that the entire population's true average height μ lies somewhere within this histogram. The histogram itself has some standard deviation σ_{mean} which is called the "standard deviation of the mean". If this σ_{mean} is very small, then our experiment was very good, and the peak of the above histogram is very close to the true μ, within a few of those σ_{mean} values.

We show below [but I never did it] that the standard deviation of the mean is given by $\sigma_{mean} = \sigma / \sqrt{N}$ where σ is the standard deviation of $p_1(x)$ shown in ***. Thus, the larger N, the smaller σ_{mean} and the better is our experiment. Notice that we don't really have to repeat the height measuring trial M times to be able to compute σ_{mean}.

Stroke Example. Again the experiment is "a person during one year" and the outcome x is a discrete binary variable which we set to x = 1 for people who had a stroke in that year, and x = 0 for those who did

not. This type of trial is sometimes called a Bernoulli Trial. Again there are N people in the trial. The outcome histogram analogous to the die histogram of ** might look like this,



Here p is the fraction of the trial people who had a stroke. From this trial we obtain an estimate $p_1(x)$ of the true probability distribution p(x) which we would find if we did this experiment for all people. The function $p_1(x)$ is very simple: $p_1(1) = p$ and $p_1(0) = 1-p$ and of course these add up to 1. It is easy to compute from this histogram our estimates μ_1 and σ_1 for mean and standard deviation:

$$\mu_1 = \Sigma_x\, p_1(x)x = (1-p)*0 + p*1 = p$$

$$\sigma_1^2 = \Sigma_x\, p_1(x)(x-\mu_1)^2 = (1-p)*(\mu_1)^2 + p*(1-\mu_1)^2 = (1-p)*(p)^2 + p*(1-p)^2$$

$$= (1-p)[\,p^2 + p(1-p)] = (1-p)[\,p] = p(1-p)$$

This is not rocket science here. Then since $\sigma_{mean} = \sigma / \sqrt{N}$, and since σ_1 is our estimate of σ, we find that

$$\sigma_{mean} = \sqrt{p(1-p)/N}$$

Now a very obscure theorem of probability called the Central Limit Theorem says that, regardless of the shape of the underlying p(x) describing the probability of some outcome x, if the experiment is carried out N times where N is reasonably large (N > 20), then the shape of the curve for the mean, like that shown in **, will be very close to a normal distribution (aka a gaussian distribution) $e^{-[(x-\mu)/(2\sigma_{mean})]^2}$. This allows one to use known facts about gaussian distributions, such as the fact that 95% of the area is contained within 1.96 standard deviations in either direction. In other words,

the true μ lies in the range $[\,\mu_1 - 1.96\,\sigma_{mean}, \mu_1 + 1.96\,\sigma_{mean}]$ with 95 % confidence

where we can compute $\sigma_{mean} \approx \sigma_1 / \sqrt{N}$ where σ_1 is our estimate for σ of the underlying p(x). In our stroke example, it happens that $\sigma_1 = p(1-p)$ so in that case $\sigma_{mean} = \sqrt{p(1-p)/N}$.

Example: Suppose N = 523 patients with a $CHADS_2$ score of 2 are studied for stroke, each for the same duration of time (maybe 1.3 years). During the study, suppose 23 of these patients had strokes, so the stroke rate is p = 23/523 = .044 = 4.4%. Since the trial did not last one year, this is not an annual rate. We then compute

$\sigma_{mean} = \sqrt{p(1-p)/N} = \sqrt{.0439(1-.0439)/524} = .00895 = .90\%$

$1.96 * \sigma_{mean} = 1.96 * .90 = 1.76\%$

$p - 1.96 * \sigma_{mean} = 4.39\% - 1.76\% = 2.64\%$

$p + 1.96 * \sigma_{mean} = 4.39\% + 1.76\% = 6.14\%$

Then we might present this result as: rate(95% CI) = 4.4% (2.6 - 6.1)

Comments: This is the path I intended for this appendix. But the actual $CHADS_2$ error bars are smaller than mine, and they are *asymmetric*. For this $CHADS_2$ group adjusted to annual basis
he gives

rate(95% CI) = 4.0% (3.1 - 5.1) 0.9 and 1.1

If I really want to understand this, I have to study more stuff. But at least I have a ballpark method which gives ballpark error bars which are reasonable. It seems that the number 523 should be large enough to have symmetric bars. But things may be different for small rates as I read somewhere.

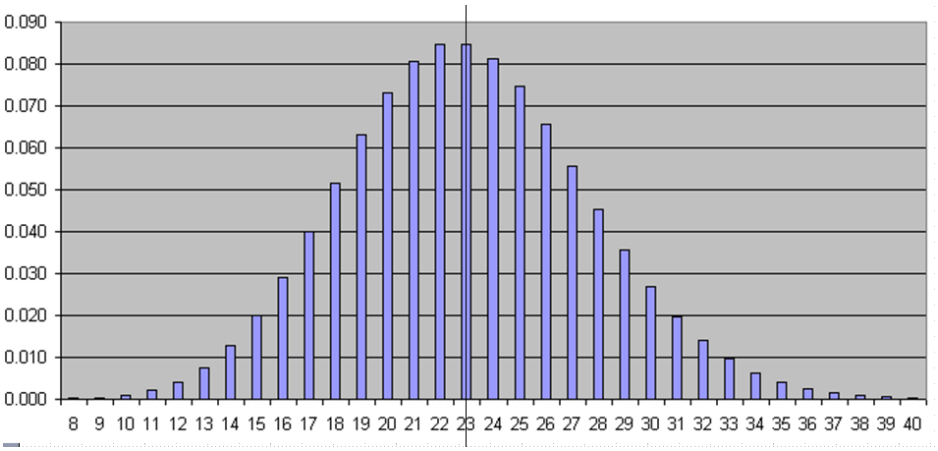
Plan A. I can see with my Excel binomial deal that when the rate is low, you don't really have a gaussian and the confidence intervals are not the same on the two sides, etc. This wiki site addresses this issue

http://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval

The stroke/no stroke really is a binomial distribution and the CLT is not applying for low rates even through N is very large.

" The central limit theorem applies poorly to this distribution with a sample size less than 30 or where the proportion is close to 0 or 1. The normal approximation fails totally when the sample proportion is exactly zero or exactly one. A frequently cited rule of thumb is that the normal approximation is a reasonable one as long as $np > 5$ and $n(1 - p) > 5$; see Brown et al. 2001.[2] "

But in my example, I have n = 523 and p = .044 and so np = 23, so this small effect should not be very significant, and when I plot this exact case with 23 out of 523 having strokes, I can see that the effect is small, The mean is a little to the right of the max, and my line is at the mean value of 23.



So I don't think this is the cause of my problem. Perhaps it is that such a test is not the same as a weighted coin toss.

Plan B. The paper says

†The adjusted stroke rate is the expected stroke rate per 100 patient-years from the exponential survival model, assuming that aspirin was not taken.

so maybe my confusion has something to do with "exponential survival model". I have read a little on this, but lets stare at the actual CHAD paper for hints.

To calculate the stroke rate as a function of $CHADS_2$, we used an exponential survival model.[27] We used the sur-

27. Lee E. *Statistical Methods for Survival Data Analysis*. 2nd ed. New York, NY: John Wiley & Sons; 1992.

Can I get this book? I found 2nd ed chapter 1. I then found 3rd ed full copy. Saved it, it is text searchable. There is an example on page 69 which is not stroke or bleed, but is "relapse" back into some kind of cancer. But I did not note there mention of an asymmetric CI. Chapter 6.1 is specific to the exponential survival model in particular. Here are the basics:

When the survival time T follows the exponential distribution with a parameter λ, the probability density function is defined as

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & t \geqslant 0, \quad \lambda > 0 \\ 0 & t < 0 \end{cases} \tag{6.1.1}$$

The cumulative distribution function is

$$F(t) = 1 - e^{-\lambda t} \qquad t \geqslant 0 \tag{6.1.2}$$

and the survivorship function is then

$$S(t) = e^{-\lambda t} \qquad t \geqslant 0 \tag{6.1.3}$$

So that, by (2.2.1), the hazard function is

$$h(t) = \lambda \qquad t \geqslant 0 \tag{6.1.4}$$

a constant, independent of t. Figure 6.1 gives the graphical presentation of the three functions.

Here F(t) is the integral of f(t) as I just checked, and F(t) and S(t) are the main acts here. λ in this model is just a parameter, same for all ages of patients I guess that means.

The mean and variance of the exponential distribution with parameter λ are, respectively, $1/\lambda$ and $1/\lambda^2$. The median is $(1/\lambda)\log 2$. The coefficient of variation is 1.

They discuss then a fancier version. How onto Example 6.1. OK, then the fancier Weibull model where there are two parameters instead of 1.

But I don't see the connection between something like an exponential survival model and a rate and its error bars.

I will try to roll my own here. You start with 523 people and you run your study for 2 years. You find that 23 people have strokes, and you know exactly when the stroke occurred for each of these people. I will continue this elsewhere.