Jericho McLeod
CSI - 873
Assigment 4

Problem 5.3

Suppose Hypothesis h commits $r = 10$ errors over a sample of $n = 65$ independently drawn examples. What is the 90% confidence interval (two-sided) for the true error rate? What is the 95% one-sided interval (i.e., what is the upper bound U such that $error_D(h) \leq U$ with 95% confidence?

The confidence interval for an observed error rate is generally expressed by:

$$errors_S(h) \pm z_N \sqrt{\frac{errors_S(h)(1 - error_S(h))}{n}}$$

In [4]:
```
import math
r = 10
n = 65
z_n_90 = 1.64

num = (r/n) * (1-(r/n))
den = n
frac = num/den
root = math.sqrt(frac)
half_range = z_n_90 * root
print('The 90% interval is',r/n,'+/-',half_range)
print('Or, from',(r/n)-half_range,'to',(r/n)+half_range)
```

```
The 90% interval is 0.15384615384615385 +/- 0.07339308747443792
Or, from 0.08045306637171594 to 0.22723924132059176
```

The 90% confidence interval implies 5% probability that the true error rate is above this range, and 5% that it is below. Thus, the upper bound from the 90% confidence interval, approximately 0.2272, is also the U below which we are 95% confident the true error rate is found.

Problem 5.4

You are about to test a hypothesis h whose $error_D$ is known to be in the range between 0.2 and 0.6. What is the minimum number of examples you must collect to assure that the width of the two-sided 95% confidence interval will be smaller than 0.1?

Using the same formula,

$$errors_S(h) \pm z_N \sqrt{\frac{errors_S(h)(1 - error_S(h))}{n}}$$

We will assume $error_D(h) = 0.4$ based on $(0.2 + 0.6)/2$, giving us

$$0.1 < \pm z_N \sqrt{\frac{0.4 * (1 - 0.4)}{n}}$$

The z score for 95% confidence is 1.96, thus:

$$0.1 < \pm 1.96 \sqrt{\frac{0.4 * (1 - 0.4)}{n}}$$

Applying the width of the confidence interval gives us:

$$0.1/(3.92) < \sqrt{\frac{0.4 * (1 - 0.4)}{n}}$$

And simplifying:

$$0.02551020408 < \sqrt{\frac{0.24}{n}}$$

$$0.00065077051 < \frac{0.24}{n}$$

$$n > \frac{0.24}{0.00065077051}$$

Thus n must be at least 368.79, and since these are tests that cannot be fractions, we can round to say we must collect at least 369 examples to ensure the width of our 95% confidence interval is smaller than 0.1.

Problem 6.1

Consider again the example application of Bayes rule in Section 6.2.1. Suppose the doctor decides to order a second laboratory test for the same patient, and suppose the second test returns a positive result as well. What are the posterior probabilies of $cancer$ and $\neg cancer$ following these two tests? Assume the two tests are independent.

Using the Bayes theorem, which states:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where A is the posterior probability and B the prior probability, the initial calculation was:

$$P(cancer|\oplus) = \frac{P(\oplus|cancer)P(cancer)}{P(\oplus)}$$

And we are given:

$P(cancer) = 0.008$
$P(\oplus|cancer) = 0.98$
$P(\ominus|cancer) = 0.02$
$P(\oplus|\neg cancer) = 0.03$
$P(\ominus|\neg cancer) = 0.97$

Then $P(\oplus|cancer)P(cancer) = .0078$
and $P(\oplus|\neg cancer)P(\neg cancer) = 0.0298$, and since these sum to 1 given the Bayes Theorem, we can normalize to 1 to determine $P(\oplus)$.

$$\frac{0.0078}{(0.0078 + 0.0298)}$$

$$\frac{0.0078}{0.0376} = 0.2074$$

Thus:

$$\frac{0.0078}{x} = 0.2074$$

With a second test returning a positive result, the are looking for the probability of cancer given two positive results:

$$P(cancer|\oplus\oplus) = \frac{P(\oplus\oplus|cancer)P(cancer)}{P(\oplus\oplus)}$$

Given that

$$P(\oplus\oplus|cancer) = P(\oplus|cancer) * P(\oplus|cancer)$$

and

$$P(\oplus\oplus) = P(\oplus\oplus|cancer) * P(cancer) + P(\oplus\oplus|\neg cancer) * P(\neg cancer)$$

which further expands to

$$P(\oplus\oplus) = P(\oplus|cancer) * P(\oplus|cancer) * P(cancer) + P(\oplus|\neg cancer) * P(\oplus|\neg cancer) * P(\neg cancer)$$

We can restate the probability being solved for as:

$$P(cancer|\oplus\oplus) = \frac{P(\oplus|cancer) * P(\oplus|cancer) * P(cancer)}{P(\oplus|cancer) * P(\oplus|cancer) * P(cancer) + P(\oplus|\neg cancer) * P(\oplus|\neg cancer) * P(\neg ca}$$

Applying this, we get:

$$P(cancer|\oplus\oplus) = \frac{0.98 * 0.98 * 0.008}{0.98 * 0.98 * 0.008 + 0.03 * 0.03 * 0.992}$$

$$P(cancer|\oplus\oplus) = \frac{0.0076832}{0.0076832 + 0.0008928}$$

$$P(cancer|\oplus\oplus) = 0.8959$$