Math3810 - Probability
Section 001 - Fall 2025
Notes: Bayes' Rule and Disease Testing

University of Colorado Denver / College of Liberal Arts and Sciences

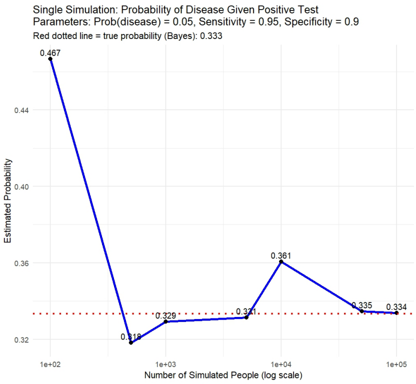
Department of Mathematics - Dr. Robert Rostermundt

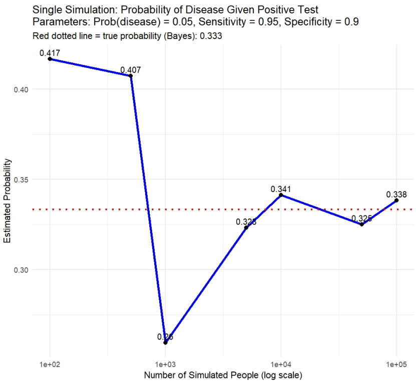
The Problem:

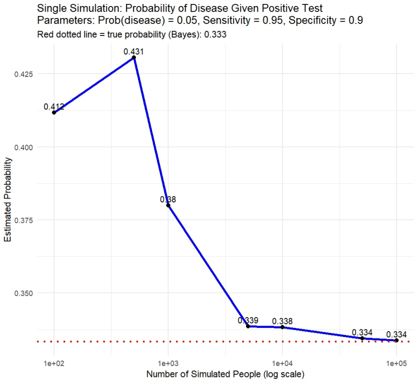
We are tested for a particular disease. We know that 5% of the population has the disease and we also have empirical data about the accuracy of the test. The test has sensitivity 95% – the ability of the test to correctly identify those who have the disease). Also, the test specificity is 90% – the ability of the test to correctly identify those who do not have the disease. We want to know the probability of having the disease given a positive test result.

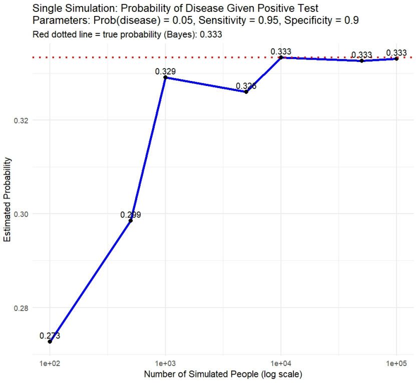
Simulations:

I have simulated this process in R. Below are graphics from four different simulations with running proportions (up to $n = 100,000$) of those having the disease from patients testing positive. In each simulation the proportion seems to be converging to a common proportion $p = 1/3$. The R code is provided the end of the document.









The following graphic shows the results of 50 different simulations. We can see that in all simulations the proportion of those having the disease given a positive test result are converging to the proportion $p = 1/3$.

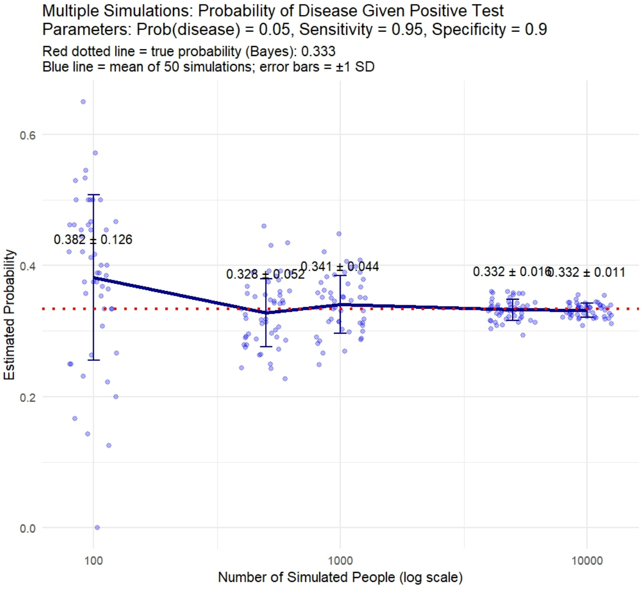


Figure 1: 50 Simulations of True Positive Test Results

Given these simulations and our frequency approach to probability theory we might suspect that the true probability is about $p = 1/3$.

Theoretical Tools:

- ***Conditional Probability:*** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $A, B \in \mathcal{F}$ where $\mathbb{P}(B) \neq 0$. Then we define the conditional probability of A given B as

$$\mathbb{P}\big(A \,|\, B\big) = \frac{\mathbb{P}\big(A \cap B\big)}{\mathbb{P}(B)}.$$

  We note that from this we can write

$$\mathbb{P}\big(A \cap B\big) = \mathbb{P}\big(B\big) \cdot \mathbb{P}\big(A \,|\, B\big) \quad \text{and} \quad \mathbb{P}\big(A \cap B\big) = \mathbb{P}\big(A\big) \cdot \mathbb{P}\big(B \,|\, A\big).$$

- ***Total Probability:*** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space where $B \in \mathcal{F}$ is a non-zero probability event. Moreover suppose $A_1, A_2, \ldots, A_n$ forms a partition of the sample space Ω where $\mathbb{P}(A_k) \neq 0$ when $1 \leq k \leq n$. Then we can write

$$\mathbb{P}\big(B\big) = \mathbb{P}\big(A_1\big) \cdot \mathbb{P}\big(B \,|\, A_1\big) + \mathbb{P}\big(A_2\big) \cdot \mathbb{P}\big(B \,|\, A_2\big) + \cdots + \mathbb{P}\big(A_n\big) \cdot \mathbb{P}\big(B \,|\, A_n\big).$$

  This is one of the main tools in a "divide and conquer" strategy in evaluating probability.

- ***Baye's Rule:*** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space where $B \in \mathcal{F}$ is a non-zero probability event. Moreover suppose $A_1, A_2, \ldots, A_n$ forms a partition of the sample space Ω where $\mathbb{P}(A_k) \neq 0$ when $1 \leq k \leq n$. Then we can write

  $$\mathbb{P}\big(A_k \,|\, B\big) = \frac{\mathbb{P}\big(A_k\big)\mathbb{P}\big(B \,|\, A_k\big)}{\displaystyle\sum_{k=1}^{n} \mathbb{P}\big(A_k\big) \cdot \mathbb{P}\big(B \,|\, A_k\big)}$$

  Bayes' Rule is simply a combination of conditional probability and total probability, but it is important when trying to make probabalistic inference. In particular, we can think of $A_1, A_2, \ldots, A_n$ as n possible states of the world and event B as evidence obtained from experiment. Then $\mathbb{P}(A_k \,|\, B)$ can be interpreted as follows: what is the probability that we are in the state A_k given the evidence outcome in event B?

Calculations:

Let D be the event that we have the disease and E be the event that we test positive for the disease. Then, since D and D^C partition the sample space, we have

$$\begin{aligned}
\mathbb{P}\big(D \,|\, E\big) &= \frac{\mathbb{P}\big(D\big) \cdot \mathbb{P}\big(E \,|\, D\big)}{\mathbb{P}\big(D\big) \cdot \mathbb{P}\big(E \,|\, D\big) + \mathbb{P}\big(D^C\big) \cdot \mathbb{P}\big(E \,|\, D^C\big)} \\
&= \frac{(0.05) \cdot (0.95)}{(0.05) \cdot (0.95) + (0.95) \cdot (0.10)} \\
&= 0.\overline{333}
\end{aligned}$$

That is, $\mathbb{P}\big(D \,|\, E\big) = 1/3$ as we suspected from the simulations. So just testing positive with a very accurate test does not mean you have a high probability of actually having the disease. With a little thought, the reason for this surprising result as the low probability of having the disease.

The R Code:

Here is the R code used for the above simulations.

```
########################################################################
# This simulates the probability of having a particular disease
# given a positive test result.

# install.packages("ggrepel") if necessary
library(ggplot2)
library(ggrepel)  # for table labels on plot

# --- Parameters ---
p_disease <- 0.05
sens <- 0.95
spec <- 0.90
```

```
# True Bayesian probability
p_true <- p_disease * sens / (p_disease * sens + (1 - p_disease) * (1 - spec))

# --- Simulation function ---
simulate_medical <- function(n, p_disease, sens, spec) {
  disease <- rbinom(n, 1, p_disease)   # 1 = has disease, 0 = does not
  test <- integer(n)

  # Conditional test results
  test[disease == 1] <- rbinom(sum(disease == 1), 1, sens)
  test[disease == 0] <- rbinom(sum(disease == 0), 1, 1 - spec)

  # Probability of disease given positive test
  p_given_positive <- mean(disease[test == 1])
  return(p_given_positive)
}

# --- Simulate multiple sample sizes ---
sample_sizes <- c(100, 500, 1000, 5000, 10000, 50000, 1e5)
results <- sapply(sample_sizes, simulate_medical,
                  p_disease = p_disease,
                  sens = sens,
                  spec = spec)

# --- Create table ---
df <- data.frame(
  n = sample_sizes,
  p_given_positive = results
)


# --- Plot ---
ggplot(df, aes(x = n, y = p_given_positive)) +
  geom_line(size = 1.2, color = "blue") +
  geom_point(size = 2) +
  scale_x_log10() +
  geom_hline(yintercept = p_true, linetype = "dotted", color = "red", size = 1) +
  geom_text(aes(label = round(p_given_positive,3)), vjust = -0.5, size = 3.5) +
  labs(
  title = paste0("Single␣Simulation:␣Probability␣of␣Disease␣Given␣Positive␣Test\n",
                      "Parameters:␣Prob(disease)␣=␣", p_disease,
                      ",␣Sensitivity␣=␣", sens,
                      ",␣Specificity␣=␣", spec),
       subtitle = paste0("Red␣dotted␣line␣=␣true␣probability␣(Bayes):␣",
       round(p_true,3)),
       x = "Number␣of␣Simulated␣People␣(log␣scale)",
       y = "Estimated␣Probability") +
  theme_minimal()


print(df)  # Table output
```

```r
################################################################################
################################################################################
# Run Multiple Simulations
################################################################################
################################################################################

library(ggplot2)
library(dplyr)

# --- Parameters ---
p_disease <- 0.05
sens <- 0.95
spec <- 0.90

# True Bayesian probability
p_true <- p_disease * sens / (p_disease * sens + (1 - p_disease) * (1 - spec))

# --- Simulation function ---
simulate_medical <- function(n, p_disease, sens, spec) {
  disease <- rbinom(n, 1, p_disease)
  test <- integer(n)
  test[disease == 1] <- rbinom(sum(disease == 1), 1, sens)
  test[disease == 0] <- rbinom(sum(disease == 0), 1, 1 - spec)
  mean(disease[test == 1])
}

# --- Multiple simulations per sample size ---
sample_sizes <- c(100, 500, 1000, 5000, 10000)
n_sims <- 50

all_results <- lapply(sample_sizes, function(n) {
  replicate(n_sims, simulate_medical(n, p_disease, sens, spec))
})

# --- Convert to long-format data frame ---
df_long <- data.frame(
  n = rep(sample_sizes, each = n_sims),
  p_given_positive = unlist(all_results)
)

# --- Summarize mean +/- SD ---
df_summary <- df_long %>%
  group_by(n) %>%
  summarize(
    mean_p = mean(p_given_positive),
    sd_p = sd(p_given_positive)
  ) %>%
  ungroup() %>%
  mutate(label = paste0(round(mean_p,3), "␣+/-␣", round(sd_p,3)))


# --- Plot ---
```

```
ggplot(df_long, aes(x = n, y = p_given_positive)) +
  geom_jitter(width = 0.1, height = 0, alpha = 0.3, color = "blue") +
  geom_line(data = df_summary, aes(x = n, y = mean_p), color = "darkblue",
  size = 1.2) +  # mean line
  geom_errorbar(data = df_summary,
                aes(x = n, ymin = mean_p - sd_p, ymax = mean_p + sd_p, y = mean_p),
                width = 0.05, color = "darkblue") +  # +/-1 SD
  geom_text(data = df_summary, aes(x = n, y = mean_p + 0.06, label = label),
            size = 3.5, color = "black") +  # table overlay above points
  scale_x_log10() +
  geom_hline(yintercept = p_true, linetype = "dotted", color = "red", size = 1) +
  labs(
    title = paste0("Multiple␣Simulations:␣Probability␣Disease␣With␣Positive␣Test\n",
                   "Parameters:␣Prob(disease)␣=␣", p_disease,
                   ",␣Sensitivity␣=␣", sens,
                   ",␣Specificity␣=␣", spec),
    subtitle = paste0(
      "Red␣dotted␣line␣=␣true␣probability␣(Bayes):␣", round(p_true,3), "\n",
      "Blue␣line␣=␣mean␣of␣", n_sims, "␣simulations;␣error␣bars␣=␣+/-1␣SD"
    ),
    x = "Number␣of␣Simulated␣People␣(log␣scale)",
    y = "Estimated␣Probability"
  ) +
  theme_minimal()


# --- Print summary table ---
print(df_summary)
```

Please let me know if you have any questions, comments, or corrections!