Homework 3

Vineet Rai

Due date: Tuesday, February 26

1. We considered a permutation test approach to compare means from two independent population in the take-home portion of the exam. In that example, we compared the observed statistic (d) with $\binom{7+16}{7} = 245,157$ many permutation statistics (d^*'s). However, permutation test quickly becomes computationally infeasible when sample sizes (n_1 and n_2) are large. In the case with large sample sizes, a natural approach is to sample a large subset of all permutation statistics and approximate the probability $P(d^* < d)$ based on that subset.

For the ease of discussion/computing, let's assume that in a separate permutation test, we have a total of permutations (500,000 d^*'s) and only 1% of these are greater than d.

a. (2 points) If 1,000 d^*'s are sampled **without replacement**, what is the probability that none of these sampled d^*'s are gerater than d?

If there are 500,000 d^* values, then 5,000 (1%) of these are greater than d and 495,000 (99%) are less than or equal to d. If we were to sample only one of these d^* values, the probability it is not greater than d is

$$P(d_1^* \leq d) = \frac{495,000}{500,000}$$

and if we were to sample two of these d^* values without replacement, the probability they are both not greater than d is

$$P(d_1^*, d_2^* \leq d) = \frac{495,000}{500,000} \times \frac{495,000-1}{500,000-1} = \frac{495,000}{500,000} \times \frac{494,999}{499,999}$$

If we sample 1,000 of these d^* values without replacement, then the probability that none of these are greater than d is

$$P(d_1^*, d_2^*, d_3^*, \cdots, d_{1000}^* \leq d) = \frac{495,000}{500,000} \times \frac{494,999}{499,999} \times \frac{494,998}{499,998} \times \cdots \times \frac{494,001}{499,001}$$

This can be rewritten as

$$P(d_1^*, d_2^*, d_3^*, \cdots, d_{1000}^* \leq d) = \frac{495,000-0}{500,000-0} \times \frac{495,000-1}{500,000-1} \times \frac{495,000-2}{500,000-2} \times \cdots \times \frac{495,000-999}{500,000-999}$$

$$= \prod_{s=0}^{999} \frac{495,000-s}{500,000-s}$$

```
> s <- 0:999
> prod((495000-s)/(500000-s)) ##computes the product of all terms
[1] 4.273722e-05
```

$$P(d_1^*, d_2^*, d_3^*, \cdots, d_{1000}^* \leq d) = 0.00004273... \approx 4.274 \times 10^{-5}$$

b. (2 points) If 1,000 d^*'s are sampled **with replacement**, what is the probability that none of these sampled d^*'s are gerater than d?

With replacement, the probability all 1,000 samples are not greater than d is

$$P(d_1^*, d_2^*, d_3^*, \cdots, d_{1000}^* \leq d) = \frac{495,000}{500,000} \times \frac{495,000}{500,000} \times \frac{495,000}{500,000} \times \cdots = 0.99^{1000}$$

```
> 0.99^1000
[1] 4.317125e-05
```

$$P(d_1^*, d_2^*, d_3^*, \cdots, d_{1000}^* \leq d) = 0.00004317... \approx 4.317 \times 10^{-5}$$

c. (2 points) Let X be a discrete random variable that represents the number of d^*'s greater than d in the sample of 1000 d^*'s in part b. Then X takes integer values $0, 1, 2, \ldots, 1000$. What is the probability distribution function of X, $f(x)$?

We have independent trials because the sampling is done with replacement. Each trial has a binary outcome of $d^* > d$ or $d^* \leq d$ for which the probabilities sum to one. Therefore, X can be modeled with a binomial probability distribution:

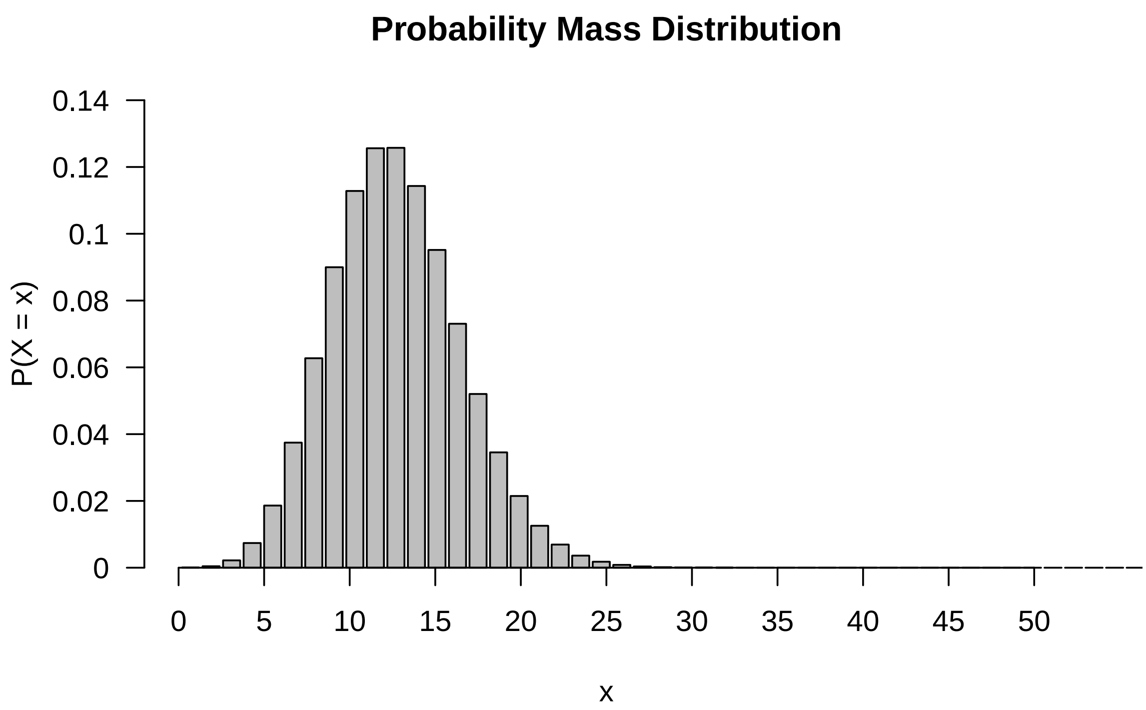
$$f(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$
$$\text{for } x = 0, 1, 2, \ldots, n$$

We have $n = 1000$ independent trials, and a success is when $d^* > d$ ($p = 0.01$) while a failure is when $d^* \leq d$ ($p = 1 - 0.01 = 0.99$). Using the formula, we have

$$f(x) = P(X = x) = \binom{1000}{x} 0.01^x 0.99^{1000-x}$$
$$\text{for } x = 0, 1, 2, \ldots, n$$

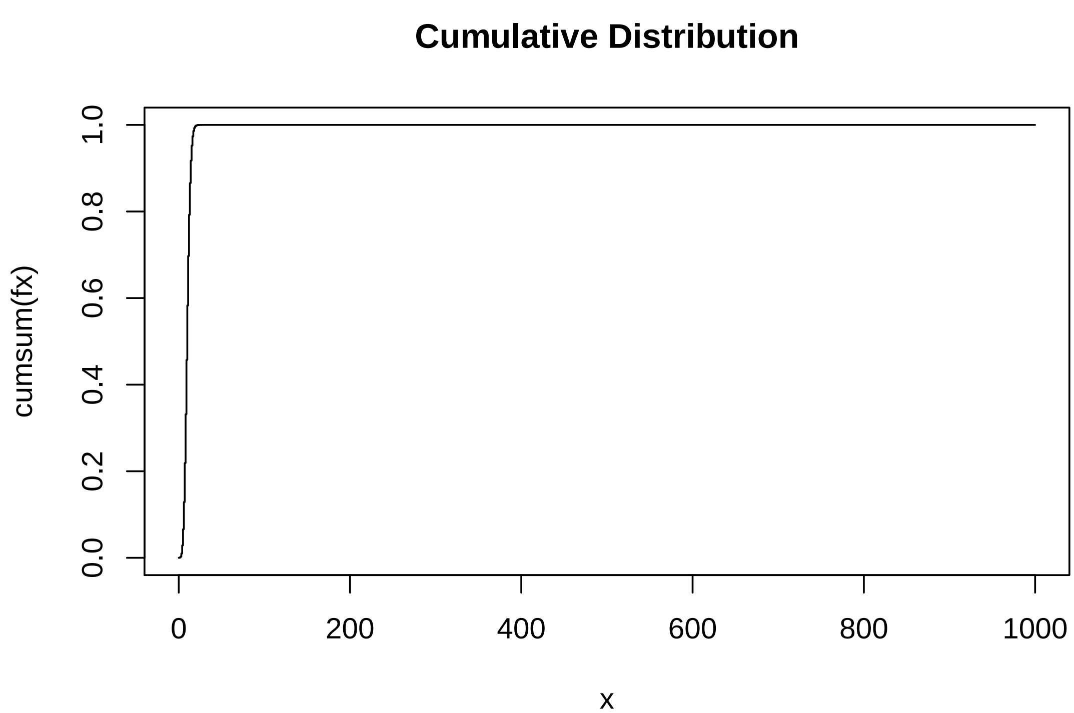
d. (2 points) Create a 'barplot' for $f(x)$ for $0 \leq X \leq 50$ (specify the range with 'xlim = c(0, 50)').

```
> x <- 0:1000
> fx <- choose(1000, x) * (0.01^x) * (1-0.01)^(1000-x)
> barplot(fx, xlim = c(0,50), ylim = c(0, 0.14),
+         main = "Probability Mass Distribution",
+         xlab = "x", ylab = "P(X = x)", axes = FALSE)
> axis(2, at = seq(0, 0.14, 0.02),
+      labels = (seq(0, 0.14, 0.02)), las = 1)
> axis(1, at = seq(0, 50, 5), labels = seq(0, 50, 5))
```



e. (2 points) Plot the cumulative distribution function of X, $F(x)$, for $0 \leq X \leq 1000$.

```
> plot(x, cumsum(fx), "s", main = "Cumulative Distribution")
```



```r
> plot(log10(x), cumsum(fx), "s", main = "Cumulative Distribution (Log Scale)")
```