

# STAT 521: Assignment 7

Make sure to show your computation and/or attach appropriate output.

## Problem 1

A researcher wishes to compare birth weights of infants among four groups of their mothers' smoking status: non-smokers, ex-smokers,  $< 1/2$  pack/day,  $\geq 1/2$  pack/day. Data is given as follows (See Assignment7.sas/R/py for computer analysis):

	Non-smoker	Ex-smoker	$< \frac{1}{2}$ pack/d	$\geq \frac{1}{2}$ pack/d	
	8.56	7.39	5.97	7.03	
	8.47	8.64	6.77	5.24	
	6.39	8.54	7.26	6.14	
	9.26	5.37	5.74	6.74	
	7.98	9.21	8.74	6.62	
	6.84	6.63	6.30	7.37	
	7.87		5.52	4.94	
				6.34	
$\sum y$	55.37	45.78	46.30	50.42	$\sum \sum y = 197.87$
$\sum y^2$	444.00	359.81	313.68	322.75	$\sum \sum y^2 = 1440.23$
$\bar{y}$	7.91	7.63	6.61	6.30	$\bar{y}_{..} = 7.07$
$n$	7	6	7	8	28

- State the null and alternative hypotheses.
- Manually calculate sum of squares and complete the ANOVA table below.

<i>Source</i>	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>
<b>Treatment</b>				
<b>Error</b>				
<b>Total</b>				

- Test the hypothesis. Use  $\alpha = 0.05$ . **What is your conclusion?**
- What are the assumptions that you are making to conduct the above test?
- Use SAS, R or Python to verify your results in part (b) and (c). Run Tukey's HSD test for multiple comparisons. Did you find any groups that were significantly different at  $\alpha = 0.05$ ? For each of significant pairs (if any), report a 95% confidence interval of the mean difference. **Attach relevant output and report your findings.**

## Problem 2

A researcher conducted a one-way ANOVA to test the null hypothesis  $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ . The researcher had a balanced design with total number of subjects  $N = 20$ . Complete the ANOVA table below.

<i>Source</i>	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>
<b>Treatment</b>		6750		
<b>Error</b>		8000		
<b>Total</b>				

## Problem 3 (Biostats students only)

*Uniform distribution:*

One of the simplest types of continuous distributions is the *uniform distribution*. The probability density function (PDF) of a uniform random variable  $X$  is defined as:

$$f(x) = \begin{cases} \frac{1}{b-a}, & a < x < b \\ 0, & \text{elsewhere} \end{cases}$$

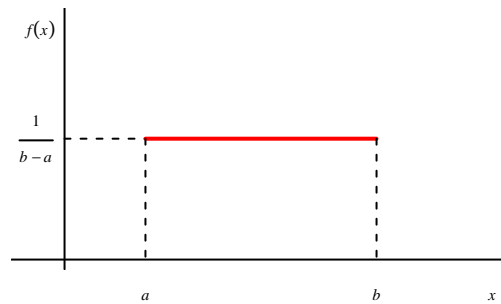


Figure 1: The probability density function of  $Uniform(a, b)$

A random variable  $X$  can take any value between  $a$  and  $b$ , and the density is constant over this range. We write this as:

$$X \sim Uniform(a, b)$$

A special case of the uniform distribution is when  $a = 0$  and  $b = 1$ :  $X \sim Uniform(0, 1)$ . This is called the *standard uniform distribution*. The standard uniform distribution has a PDF of  $f(x) = 1$  for  $0 < x < 1$  and  $f(x) = 0$  otherwise.

**Question:** Use the definition of the expected value  $E(X) = \int_{-\infty}^{\infty} x f(x) dx$  (see Assignment #3) and show that  $E(X) = \frac{a+b}{2}$  and  $var(X) = \frac{(b-a)^2}{12}$  if  $X \sim Uniform(a, b)$ .

#### Problem 4 (Biostats students only)

Monte Carlo integration using R

Suppose we wish to solve a definite integral  $I = \int_a^b g(x) dx$ ,  $a < b$ . Note that  $g(x)$  is any one-dimensional function, not necessarily a PDF. Sometimes  $g(x)$  is very complex and there may not be a closed-form solution. We can rewrite the integral as:

$$I = \int_a^b g(x) dx = \int_a^b \frac{g(x)}{f(x)} f(x) dx$$

If we can find a random variable that has a well-known PDF of  $f(x)$  whose support is  $a < x < b$ , then by the definition of the expected value, the integral is equivalent to:

$$I = \int_a^b \frac{g(x)}{f(x)} f(x) dx = E \left[ \frac{g(x)}{f(x)} \right]$$

in which a random variable  $X$  has a PDF  $f(x) \geq 0$  for  $a < x < b$  and  $f(x) = 0$  otherwise. An obvious choice of  $X$  is a uniform random variable  $X \sim \text{Uniform}(a, b)$  with  $f(x) = \frac{1}{b-a}$  for  $a < x < b$ . Then it follows:

$$I = E \left[ \frac{g(x)}{f(x)} \right] = E \left[ \frac{g(x)}{1/(b-a)} \right] = (b-a)E[g(x)]$$

What does this imply? Remember, from Assignment #6, you can approximate  $E[g(x)]$  as long as you can generate a large number of random values from  $X \sim \text{Uniform}(a, b)$ . If we take the average of all the random values of  $g(x)$  and multiply by  $(b-a)$ , we can approximate the definite integral  $I = \int_a^b g(x) dx$ .

**Example:** Suppose we want to solve:

$$I = \int_0^2 \frac{2e^{-2x}}{(1+e^{-2x})^2} dx$$

Let  $g(x) = \frac{2e^{-2x}}{(1+e^{-2x})^2}$  and  $X$  be a uniform random variable  $X \sim \text{Uniform}(0, 2)$  whose PDF is  $f(x) = \frac{1}{2}$  for  $0 < x < 2$ . Then we can approximate the integral  $I$  as  $2E[g(X)]$ .

Now let's do this with R. First I'm going to produce one million random numbers from  $X \sim \text{Uniform}(0, 2)$ . This can be done using `runif()` function. Its syntax is `runif(n, a, b)` where `n` is the number of random values you want to generate, and `a` and `b` are the range of  $x$ . In the code below, I assign values  $a = 0$  and  $b = 2$ .

```
# Generate 1 million random number from X ~ Unif(0, 2)
a <- 0
b <- 2
x <- runif(10 ^ 6, a, b)
head(x)
```

```
## [1] 0.5751550 1.5766103 0.8179538 1.7660348 1.8809346 0.0911130
```

Then for each value of  $x$ , I calculate  $g(x)$ . For example, for the first uniform random value of 0.575155, we have  $g(0.575155) = \frac{2e^{-2(0.575155)}}{(1+e^{-2(0.575155)})^2} \approx 0.365$ . For  $e^x$ , use `exp()` function.

```
# Calculate g(x)
g <- 2 * exp(-2 * x) / (1 + exp(-2 * x)) ^ 2
head(g)
```

```
## [1] 0.36524937 0.07857294 0.27289256 0.05521215 0.04439313 0.49587208
```

Now, to approximate  $I = (b - a)E[g(x)]$ , we calculate the mean of `g` and then multiply by  $b - a = 2$ .

```
# Take the average and multiply by (b - a)  
(b - a) * mean(g)
```

```
## [1] 0.4825734
```

In this case, I got  $I \approx 0.4825734$ . This is close to the correct answer of  $I = \int_0^2 \frac{2e^{-2x}}{(1+e^{-2x})^2} dx = 0.4820138$ .

**Here's an exercise for you:**

Approximate:

$$I = \int_{0.2}^{0.3} 15(3x)^4 e^{-(3x^5)} dx$$

Please submit your R code as well.