Module 6 Peer Review Assignment

Problem 1

Suppose X and Y are independent normal random variables with the same mean μ and the same variance σ^2. Do the random variables $W = X + Y$ and $U = 2X$ have the same distribution? Explain.

$$E[W] = E[X] + E[Y] = \mu + \mu = 2\mu$$
$$Var(W) = Var(X) + Var(Y) = \sigma^2 + \sigma^2 = 2\sigma^2, \text{ since } X, Y \text{ independent}$$
$$\implies W \sim \mathcal{N}(2\mu, 2\sigma^2)$$
$$E[U] = E[2X] = 2\mu$$
$$Var(U) = Var(2X) = 2^2 . Var(X) = 4\sigma^2$$
$$\implies U \sim \mathcal{N}(2\mu, 4\sigma^2)$$

Hence, the distributions are not the same.

Problem 2: Central Limit Theorem and Simulation

a) For this problem, we will be sampling from the Uniform distribution with bounds [0, 100]. Before we simulate anything, let's make sure we understand what values to expect. If $X \sim U(0, 100)$, what is $E[X]$ and $Var(X)$?

$$E[X] = \frac{0+100}{2} = 50$$

$$Var(X) = \frac{(100-0)^2}{12} = 833.33$$

b) In real life, if we want to estimate the mean of a population, we have to draw a sample from that population and compute the sample mean. The important questions we have to ask are things like:

- Is the sample mean a good approximation of the population mean?
- How large does my sample need to be in order for the sample mean to well-approximate the population mean?

Complete the following function to sample n rows from the $U(0, 100)$ distribution and return the sample mean. Start with a sample size of 10 and draw a sample mean from your function. Is the estimated mean a good approximation for the population mean we computed above? What if you increase the sample size?

In [2]:
```
uniform.sample.mean = function(n){

    # Your Code Here
    sample = runif(n, 0, 100)
    sample.mean = mean(sample)
    return(sample.mean)
}

uniform.sample.mean(10)
```

53.294711294584

Since sample size is small, it's not a very good approximation of population mean, since the error in approximation is $\approx \frac{3.3}{50} \approx 6.6\%$.

c) Notice, for a sample size of n, our function is returning an estimator of the form

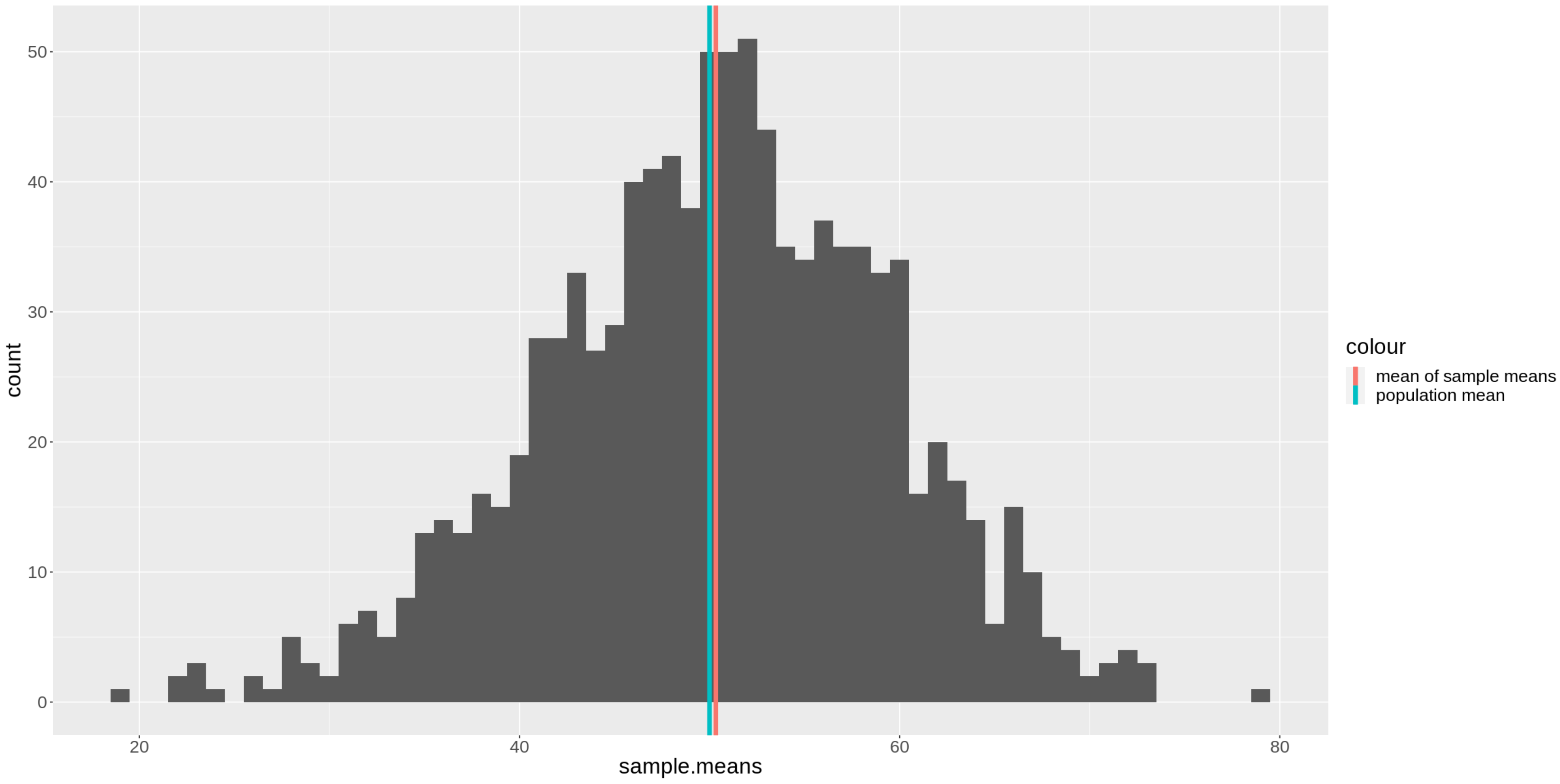
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

That means, if each X_i is a random variable, then our sample mean is also a random variable with its own distribution. We call this disribution the sample distribution. Let's take a look at what this distribution looks like.

Using the `uniform.sample.mean` function, simulate $m = 1000$ sample means, each from a sample of size $n = 10$. Create a histogram of these sample means. Then increase the value of n and plot the histogram of those sample means. What do you notice about the distribution of $\bar{X}$? What is the mean μ and variance σ^2 of the sample distribution?

In [25]:
```
# Your Code Here
options(repr.plot.width=20, repr.plot.height=10)
set.seed(2)
m <- 1000
n <- 10
mu <- 50
var <- 833.333
sample.means <- replicate(m, uniform.sample.mean(n))
mean <- mean(sample.means)
var <- var(sample.means)
print(c(mean, var))
library(ggplot2)
ggplot() + geom_histogram(aes(sample.means), binwidth=1) +
           geom_vline(aes(xintercept=mean, color='mean of sample means'), lwd=2) +
           geom_vline(aes(xintercept=mu, color='population mean'), lwd=2) +
           theme(text = element_text(size = 20))
```

```
[1] 50.33065 83.23977
```



Population mean $= \mu$, variance $= \sigma^2$

By the CLT,

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{D} \mathcal{N}(0, 1)$$

or

$$\bar{X} \xrightarrow{D} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \text{ as } n \to \infty$$

BY SLLN,

$$\bar{X} \xrightarrow{a.s.} \mu \text{ as } n \to \infty$$

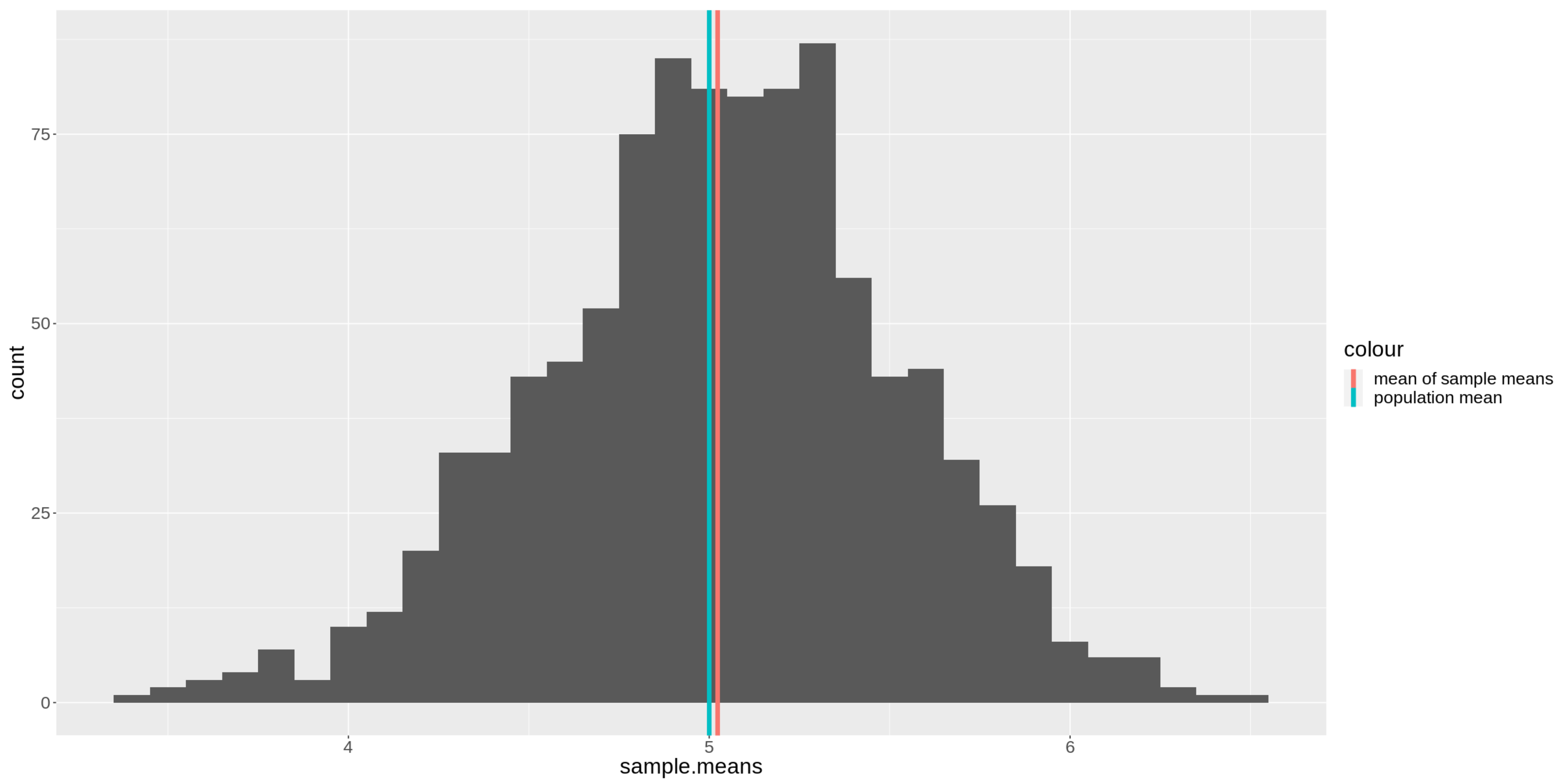
Hence with large sample size n, mean of the sampling distribution converges to population mean $\mu = 50$, with variance of the distribution as $\frac{\sigma^2}{n} = \frac{833.33}{10} = 83.33$, i.e., standard error of sample mean $= SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ and the mean of the sampling distribution converges (in distribution) to $\mathcal{N}(50, 83.33)$, as can be seen from the above histogram.

d) Recall that our underlying population distribution is $U(0, 100)$. Try changing the underlying distribution (For example a binomial(10, 0.5)) and check the sample distribution. Be sure to explain what you notice.

In [29]:
```
# Your Code Here
binom.sample.mean = function(n){

    # Your Code Here
    sample = rbinom(n, 10, 0.5)
    sample.mean = mean(sample)
    return(sample.mean)
}

binom.sample.mean(10)

set.seed(2)
m <- 1000
n <- 10
mu <- 10*0.5
var <- 10*0.5*(1-0.5)
sample.means <- replicate(m, binom.sample.mean(n))
mean <- mean(sample.means)
var <- var(sample.means)
print(c(mean, var))
library(ggplot2)
ggplot() + geom_histogram(aes(sample.means), binwidth=0.1) +
           geom_vline(aes(xintercept=mean, color='mean of sample means'), lwd=2) +
           geom_vline(aes(xintercept=mu, color='population mean'), lwd=2) +
           theme(text = element_text(size = 20))
```

4.6

```
[1] 5.0233000 0.2459731
```



For $X_i \sim B(10, 0.5)$, we have population mean $\mu = 10 * 0.5 = 5$ and variance $= 10 * 0.5 * (1 - 0.5) = 2.5$

Again, with large sample size n, mean of the sampling distribution converges to population mean $\mu = 5$ (by SLLN, CLT), with variance of the distribution as $\frac{\sigma^2}{n} = \frac{2.5}{10} = 0.25$ and the mean of the sampling distribution converges (in distribution) to $\mathcal{N}(5, 0.25)$, as can be seen from above.

Hence, we can see if n samples (random varibles $X_i, \ i = 1, 2, \ldots n$) are i.i.d. (independently identically distributed r.v.s, drawn from the population with mean μ, s.d. σ), then the mean of the sampling distribution $\bar{X}$ converges in distribtion to $\mathcal{N}(\mu, \frac{\sigma^2}{n})$, no matter what the underlying distribution of the variables is (by the CLT).

Problem 3

Let X be a random variable for the face value of a fair d-sided die after a single roll. X follows a discrete uniform distribution of the form $\text{unif}\{1, d\}$. Below is the mean and variance of $\text{unif}\{1, d\}$.

$$E[X] = \frac{1+d}{2} \qquad Var(X) = \frac{(d-1+1)^2 - 1}{12}$$

a) Let $\bar{X}_n$ be the random variable for the mean of n die rolls. Based on the Central Limit Theorem, what distribution does $\bar{X}_n$ follow when $d = 6$.

By the CLT, $X_n \sim \mathcal{N}\left(E[X], \frac{Var(X)}{n}\right)$

For $d = 6$, $E[X] = \frac{7}{2}$ and $Var(X) = \frac{35}{12}$, hence we have $X_n \sim \mathcal{N}\left(\frac{7}{2}, \frac{35}{12n}\right)$

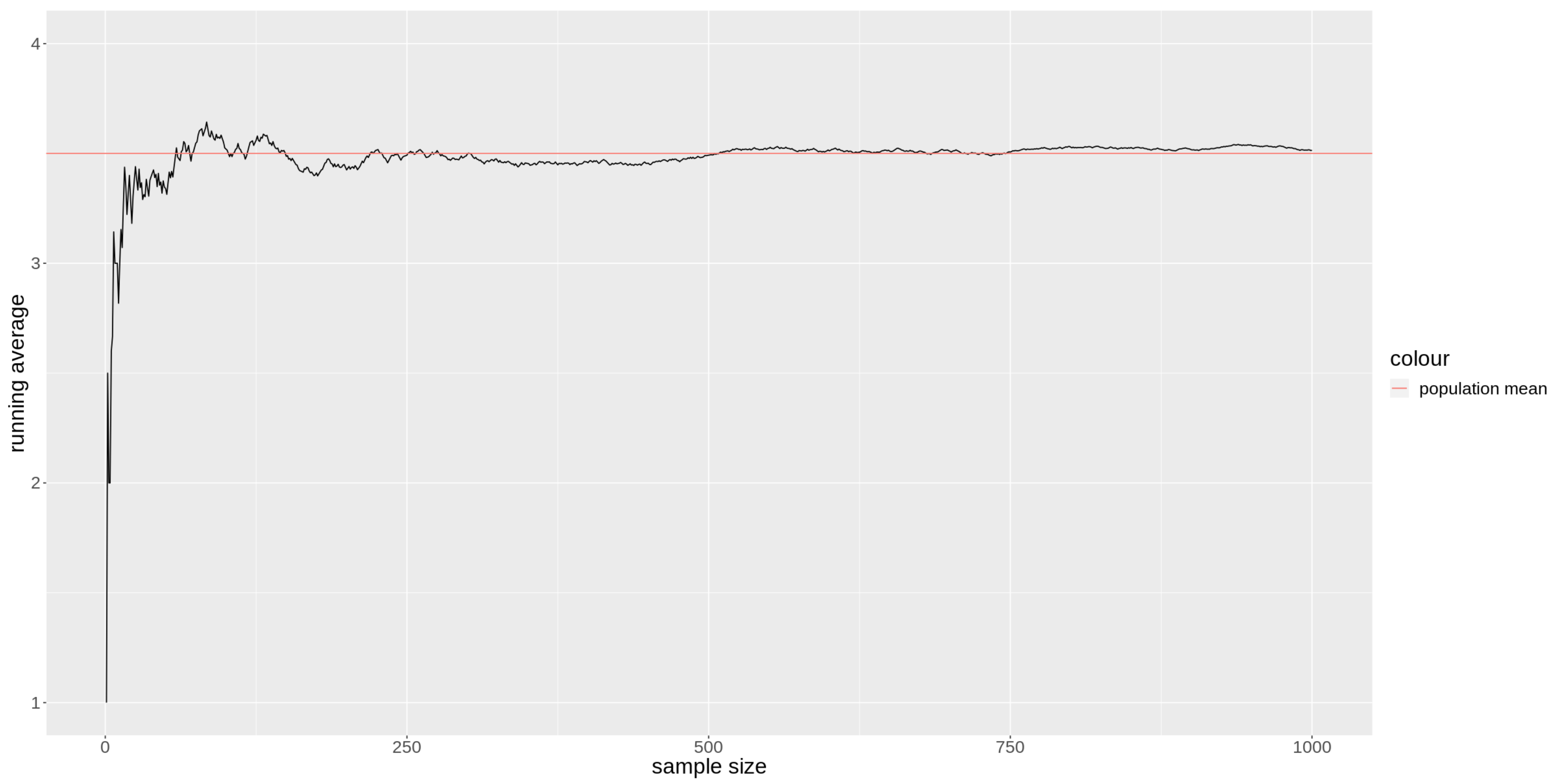
b) Generate $n = 1000$ die values, with $d = 6$. Calculate the running average of your die rolls. In other words, create an array r such that:

$$r[j] = \sum_{i=1}^{j} \frac{X_i}{j}$$

Finally, plot your running average per the number of iterations. What do you notice?

In [40]:
```
# Your Code Here
set.seed(1)
n = 1000
X = rep(0, n)
r = rep(0, n)
for (j in 1:n) {
    X[j] = sample(1:6, 1, replace=TRUE)
    r[j] = sum(X[1:j])/j
}
#c(mean(X), var(X/n), 35/12/n)
ggplot() + geom_line(aes(1:n, r)) + xlab('sample size') + ylab('running average') + ylim(1,4) +
           geom_hline(aes(yintercept=7/2, color='population mean')) +
           theme(text = element_text(size = 20))
```

3.514 · 2.93674074074074e-06 · 0.00291666666666667



As we can see from the above running averge plot, the mean of the sampling distribution converges to population mean $\frac{7}{2}$, as the sample size increases (by LLN).