Chapter 4

Similar to last chapter, we can calculate probabilities for common continuous random variables in R and generate random data according to continuous distributions.

R uses four prefixes to reference difference elements of a random variable. These are:

p for "probability", the cumulative distribution function (CDF)
q for "quantile", the inverse CDF
d for "density", the probability mass function (PMF)
r for "random", a random variable having the specified distribution

Suffixes for continuous random variables include: **norm** (normal/Gaussian), **exp** (exponential), and **t** (t).

Suppose $X \sim \mathcal{N}(\mu = 100, \sigma = 15)$, the distribution often used to model IQ scores. Suppose we want to know the probability of a randomly chosen IQ score below 110.

```
pnorm(q=110,mean=100,sd=15)

## [1] 0.7475075
```

In other words, about 75% of IQ scores are less than 110.

Let's calculate the probability that a randomly chosen score is above 130. Note that $P(X > 130) = 1 - P(X \leq 130)$.

```
1-pnorm(q=130,mean=100,sd=15)

## [1] 0.02275013
```

In other words, about 2.3% of IQ scores are greater than 130.

We can also do inverse normal calculations using the prefix **q** which stands for quantile. Suppose we want to know the 90th percentile of IQ scores. We can calculate this using the **qnorm** function.

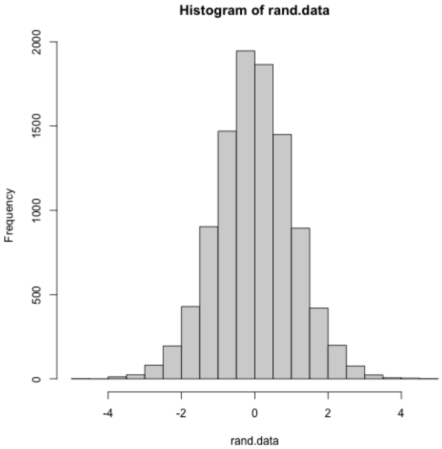
```
qnorm(p=0.9,mean=100,sd=15)

## [1] 119.2233
```

This result shows that the 90th percentile of IQ scores corresponds to an approximate IQ of 119.

In the second half of the course, we will be using the **t-distribution** frequently when dealing with statistical applications. Let's generate 10,000 random values for a t-distribution with 20 degrees of freedom, i.e., $Z \sim t(df = 20)$.

```
set.seed(2020)
rand.data = rt(n=10000,df=20)
hist(rand.data)
```



Based on the plot above, this distribution is approximately symmetric, bell-shaped, and unimodal.

Let's calculate some summary statistics for this data.

```
mean(rand.data)

## [1] -0.009077007

median(rand.data)

## [1] -0.01514699
```

The mean and median are both approximately equal to zero and since these values are nearly equal, we can confirm that the distribution is approximately symmetric.
Finally, let's calculate the probability in our simulated data so that $P(Z > 1)$.

```
num.events = sum(rand.data > 1)
prob = num.events / 10000; prob

## [1] 0.1623
```

Approximately 16% of our randomly generated t-distribution values are greater than 1.