

American Community Survey Exercise:

1. What are the elements in your data (including the categories and data types)?

```
'data.frame':      136 obs. of  8 variables:
 $ Id                : chr
 $ Id2               : integer
 $ Geography         : chr
 $ PopGroupID        : integer
 $ POPGROUP.display.label: chr
 $ RacesReported     : integer
 $ HSDegree           : number
 $ BachDegree        : number
```

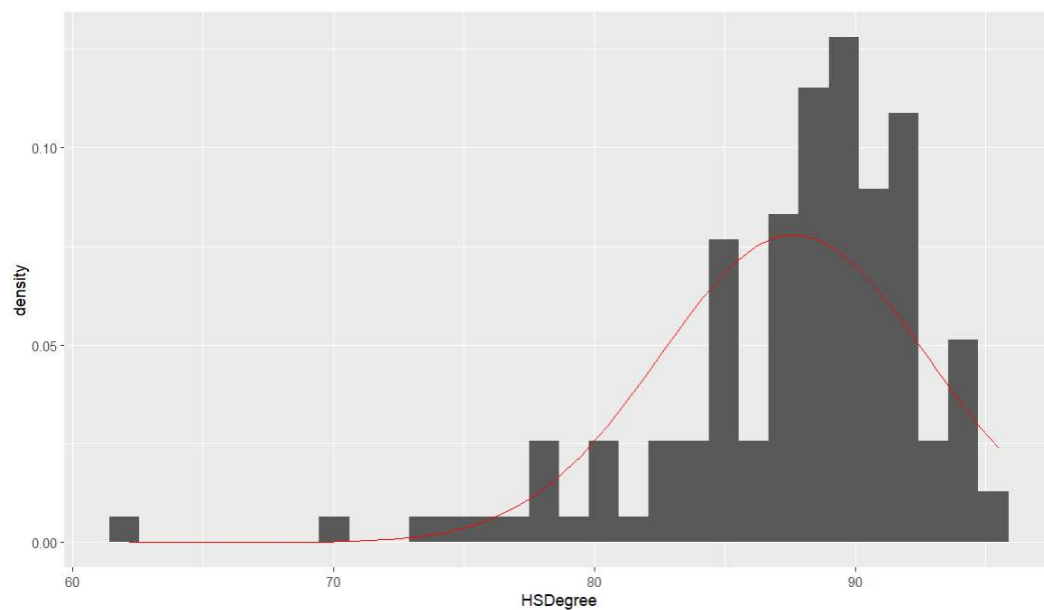


Figure 1: high school degree holder percentage per population in different counties

2. Histogram observation

- This data distribution is unimodal.
- The distribution is not symmetrical.
- The distribution is close to a bell shape but has multiple high peaks with concentrated bell shape.
- The distribution is not normal.
- The distribution is negatively skewed.
- A normal distribution is not preferred to model this data.

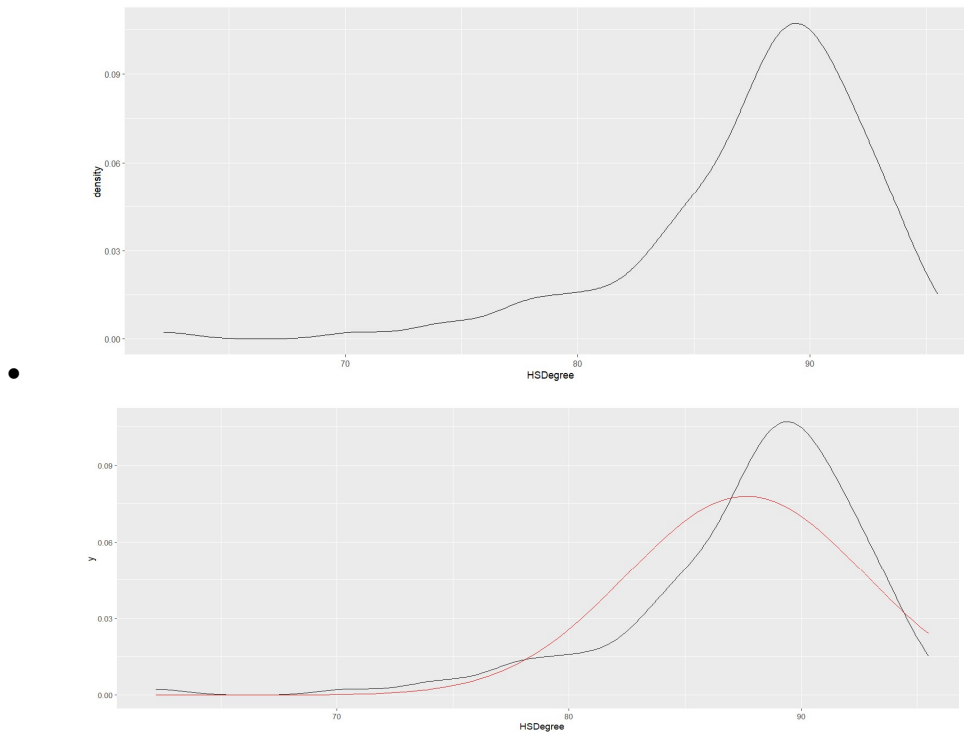


Figure 2: probability chart: high school degree holder percentage per population in different counties

probability observation

- This distribution is not normal, since it has peak not centered.
- It is negatively skewed toward right side. The left side has long tail but right side has sharp and short decline.

```
> stat.desc(data$HSDegree, basic = TRUE, desc = TRUE, norm = TRUE, p = 0.95)
```

nbr.val	nbr.null	nbr.na	min	max	range	sum	median	mean	SE.mean
1.360000e+02	0.000000e+00	0.000000e+00	6.220000e+01	9.550000e+01	3.330000e+01	1.191800e+04	8.870000e+01	8.763235e+01	4.388598e-01
CI.mean.0.95	var	std.dev	coef.var	skewness	skew.2SE	kurtosis	kurt.2SE	normtest.W	normtest.p
8.679296e-01	2.619332e+01	5.117941e+00	5.840241e-02	-1.674767e+00	-4.030254e+00	4.352856e+00	5.273885e+00	8.773635e-01	3.193634e-09

Summary:

- The skewness = -1.67, since it is less than -1 , it is highly skewed and skewed toward left. Which is proven by chart above.
- The kurtosis = 4.35, a normal distribution should get 3. So it have thin bell.
- Since the kurt.2SE = 5.27 and skew.2SE = -4.03. so the skew is unlikely by chance. It is a real skewness.
- Since Z score is normalized, this analysis works for small samples size, if the sample size is large, it won't be accurate and need to look at the actual distribution shape.