* Distribution of the Sample Mean

When we take a random sample and calculate the mean, that sample mean is a random variable as well. It depends on the random sample we took.

Let $\bar{Y}$ denote the r.v. of all possible sample means, when we sample from a population with mean μ_Y, std. dev σ_Y. Any sample mean we measure is $\bar{y}$, which is an observed value of $\bar{Y}$. In general, $\bar{Y} = \frac{1}{n}\sum Y_i$, Y_i's random.

Some Facts about $\bar{Y}$:

i) $\mu_{\bar{Y}} = \mu_Y$ (The average value of the sample mean is μ_Y)

ii) $\sigma_{\bar{Y}} = \sigma_Y/\sqrt{n}$ (The spread for the sample mean is less than σ_Y)

Notice if our "population" were: 2, 4, 6, 8 and we sampled without replacement, if $n = 2$ The two smallest values are (2,4), which have a mean of 3. The two highest are (6,8), which have a mean of 7.

So, the min and max of the population are 2 and 8, but the min and max of $\bar{Y}$ when $n = 2$ are 3 and 7. Thus, the spread shrinks.

$\bar{Y}$ and the normal distribution:

There are two results that we will use often:

I) If the population Y is normal, $\bar{Y}$ is also normal no matter what n is. I.e, if $Y \sim N(\mu_Y, \sigma_Y)$, then $\bar{Y} \sim N(\mu_Y, \sigma_{\bar{Y}} = \sigma_Y/\sqrt{n})$ for any n.

II) * The Central Limit Theorem: If we randomly sample $n \geq 30$ subjects from a population with mean μ_Y, std. dev. σ_Y, then $\bar{Y} \sim N(\mu_Y, \sigma_{\bar{Y}} = \sigma_Y/\sqrt{n})$ no matter what the population distribution

We use I) and II) to standardize $\bar{Y}$ to Z:

$$Z = (\bar{Y} - \mu_{\bar{Y}})/\sigma_{\bar{Y}} = (\bar{Y} - \mu_Y)/(\sigma_Y/\sqrt{n})$$

Ex: The heights of a certain breed of corn plant are normally distributed with mean 145cm, std dev 22cm.

a) If we randomly sample 16 corn plants, what is the prob. that their average is over 155cm?

Since $Y \sim N(145, \sigma_Y = 22)$, $\bar{Y} \sim N(145, \sigma_{\bar{Y}} = 22/\sqrt{16})$

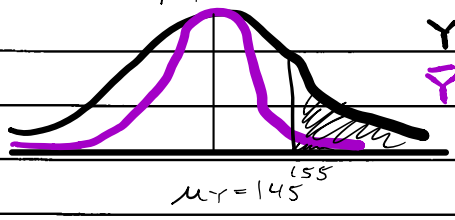
$$P_r\{\bar{Y} > 155\} = P_r\left\{Z > \frac{155-145}{(22/\sqrt{16})}\right\} = P_r\{Z > 1.82\}$$
$$= 1 - P_r\{Z < 1.82\} = 1 - 0.9656 = 0.0344$$

b) If we randomly sample 1 corn plant, is it more or less likely to be over 155cm tall compared to the sample mean of 16 corn plants?

$$P_r\{Y > 155\} = P_r\left\{Z > \frac{155-145}{22}\right\} = P_r\{Z > 0.45\} = 1 - P_r\{Z < .45\}$$
$$= 1 - 0.6736 = 0.3264$$

More likely, since the prob is higher.

OR



The variance of a single plant is higher, and the mean of Y and $\bar{Y}$ is the same, so the prob of being high values is larger for Y.

c) Find the 60th percentile for the average height of 16 randomly sampled corn plants.

$Z^{(60)} = 0.25$ (.5987 is in row 0.2, column 0.05 and is closest to 0.60)

Then, since $Z^{(60)} = \dfrac{\bar{Y}^{(60)} - \mu_{\bar{Y}}}{\sigma_{\bar{Y}}}$,

$$\bar{Y}^{(60)} = \mu_{\bar{Y}} + Z^{(60)}\sigma_{\bar{Y}} = \mu_Y + Z^{(60)}\frac{\sigma_Y}{\sqrt{n}} = 145 + 0.25\left(\frac{22}{\sqrt{16}}\right)$$
$$= 146.375$$

* Asessing Normality

Since normality allows for easy manipulation of otherwise difficult data, we often want to check to see if data is normal. Here are two ways:

1) The empirical rule: If your data is approximately normal,
~68% of the data should lie between $(\bar{y}-s, \bar{y}+s)$
~95% of the data should lie between $(\bar{y}-2s, \bar{y}+2s)$
~99.7% of the data should lie between $(\bar{y}-3s, \bar{y}+3s)$
Note that $(\bar{y}-(k)s, \bar{y}+k(s))$ is simply k std dev's from the mean.

Note: If the data was perfectly normal,
$P_r\{\mu_Y-\sigma_Y < Y < \mu_Y+\sigma_Y\} = P_r\{\frac{(\mu_Y-\sigma_Y)-\mu_Y}{\sigma_Y} < Z < \frac{(\mu_Y+\sigma_Y)-\mu_Y}{\sigma_Y}\}$
$= P_r\{-1 < Z < 1\} = P_r\{Z<1\} - P_r\{Z<-1\}$
$= 0.8413 - 0.1587 = 0.6826$ or 68.26%

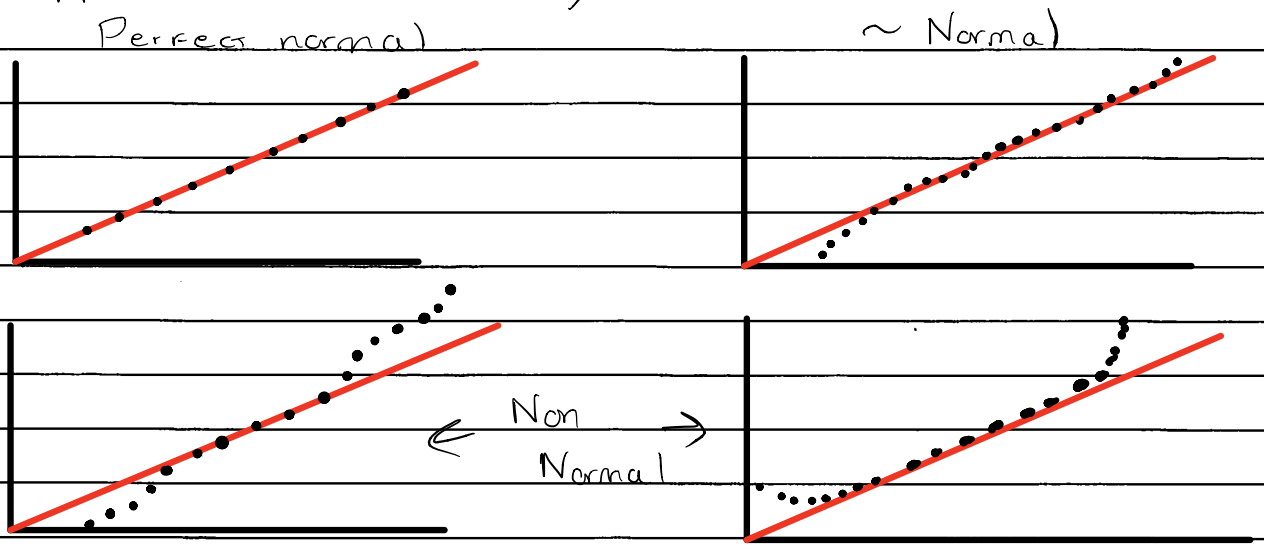
$P_r\{\mu_Y-2\sigma_Y < Y < \mu_Y+2\sigma_Y\} = P_r\{-2 < Z < 2\}$
$= P_r\{Z<2\} - P_r\{Z<-2\} = 0.9772 - 0.0228 = 0.9544$ or 95.44%

$P_r\{\mu_Y-3\sigma_Y < Y < \mu_Y+3\sigma_Y\} = P_r\{-3 < Z < 3\}$
$= P_r\{Z<3\} - P_r\{Z<-3\} = 0.9987 - 0.0013 = 0.9974$ or 99.74%

So using our sample data, you would calculate $(\bar{y}-s, \bar{y}+s)$, $(\bar{y}-2s, \bar{y}+2s)$, $(\bar{y}-3s, \bar{y}+3s)$, and calculate the relative freq of points that fall in each, and compare that to ~68%, ~95%, ~99.7%.

2) QQ plots / Normal Probability plots
A QQ plot calculates the percentiles for all y_i's, their z-scores, and their theoretical percentiles based on their z-scores (uses the normal distribution). Then, it plots (actual percentiles) vs (normal percentiles).

If your data was perfectly normal, the QQ plot would be the line $y = x$. Typically we look for "approximate" normality



Note: These plots are subjective, but in general you should look at the proportion of points that fall away from the line (or significantly away from the line) in order to asess normality.