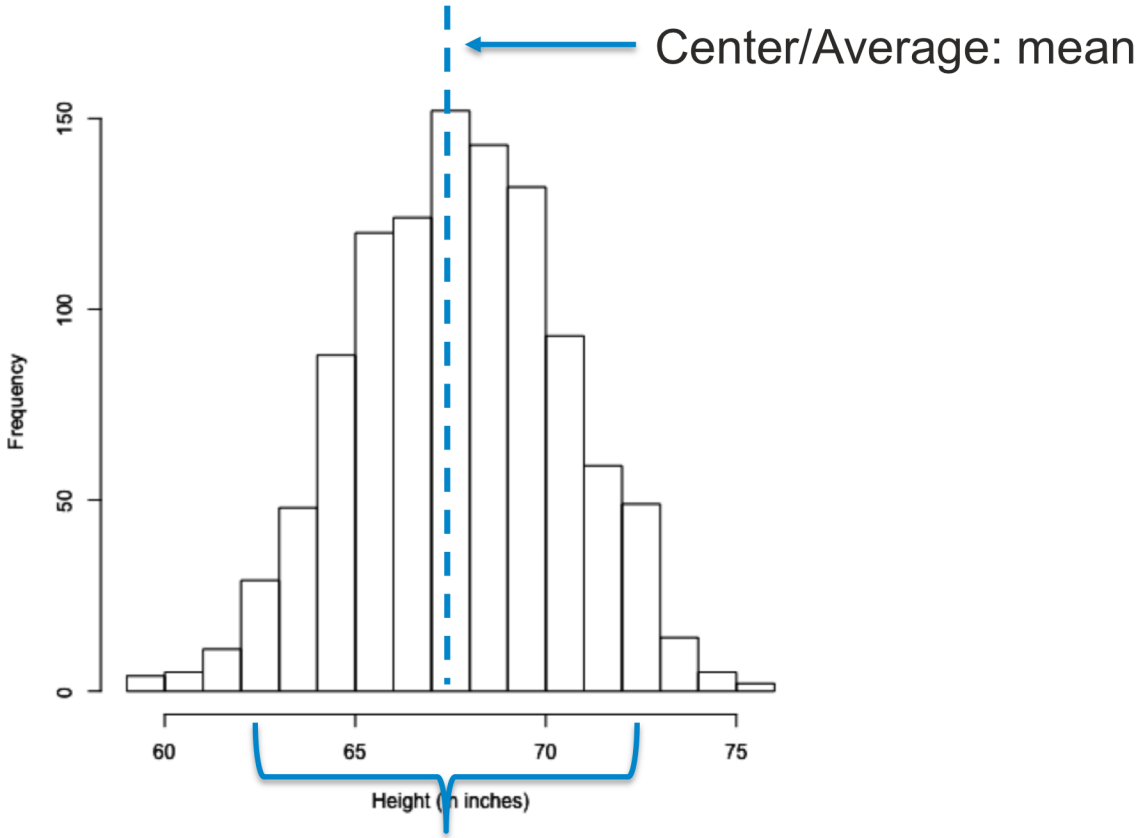
Populations, Samples, and the Central Limit Theorem

Means, Variances, and Standard Deviations



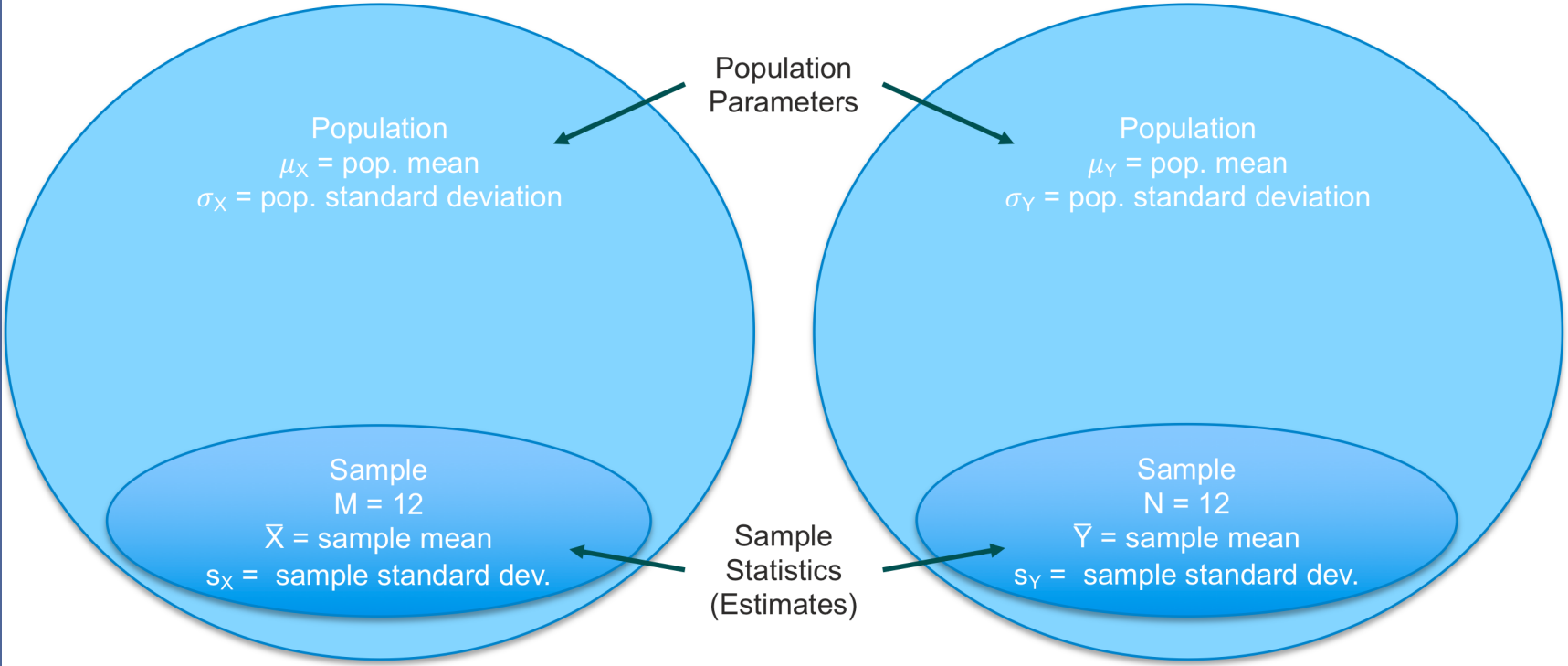
Variation: variance and standard deviation

What is a population?

- First step: understand/ID/define population(s)
  - It's what you're interested in
  - Sometimes it's theoretical
- Our mouse bodyweight example – two populations:
  - Control
  - High fat diet

Control Population: all female mice on control diet

High-Fat (HF) Population: all female mice on high-fat diet



$$\bar{Y} - \bar{X} \sim \mu_Y - \mu_X$$

Means, Variances, and Standard Deviations

Population Parameters (fixed):

$$\mu_X = \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \sigma_X^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_X)^2 \qquad \sigma = \sqrt{\sigma^2}$$

Sample Statistics (random variables):

$$\bar{X} = \frac{1}{M}\sum_{i=1}^{M} X_i = \frac{X_1 + X_2 + X_3 + \cdots + X_M}{M}$$

$$s_X^2 = \frac{1}{M-1}\sum_{i=1}^{M}(X_i - \bar{X})^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_M - \bar{X})^2}{M - 1} \qquad s_X = \sqrt{s_X^2}$$

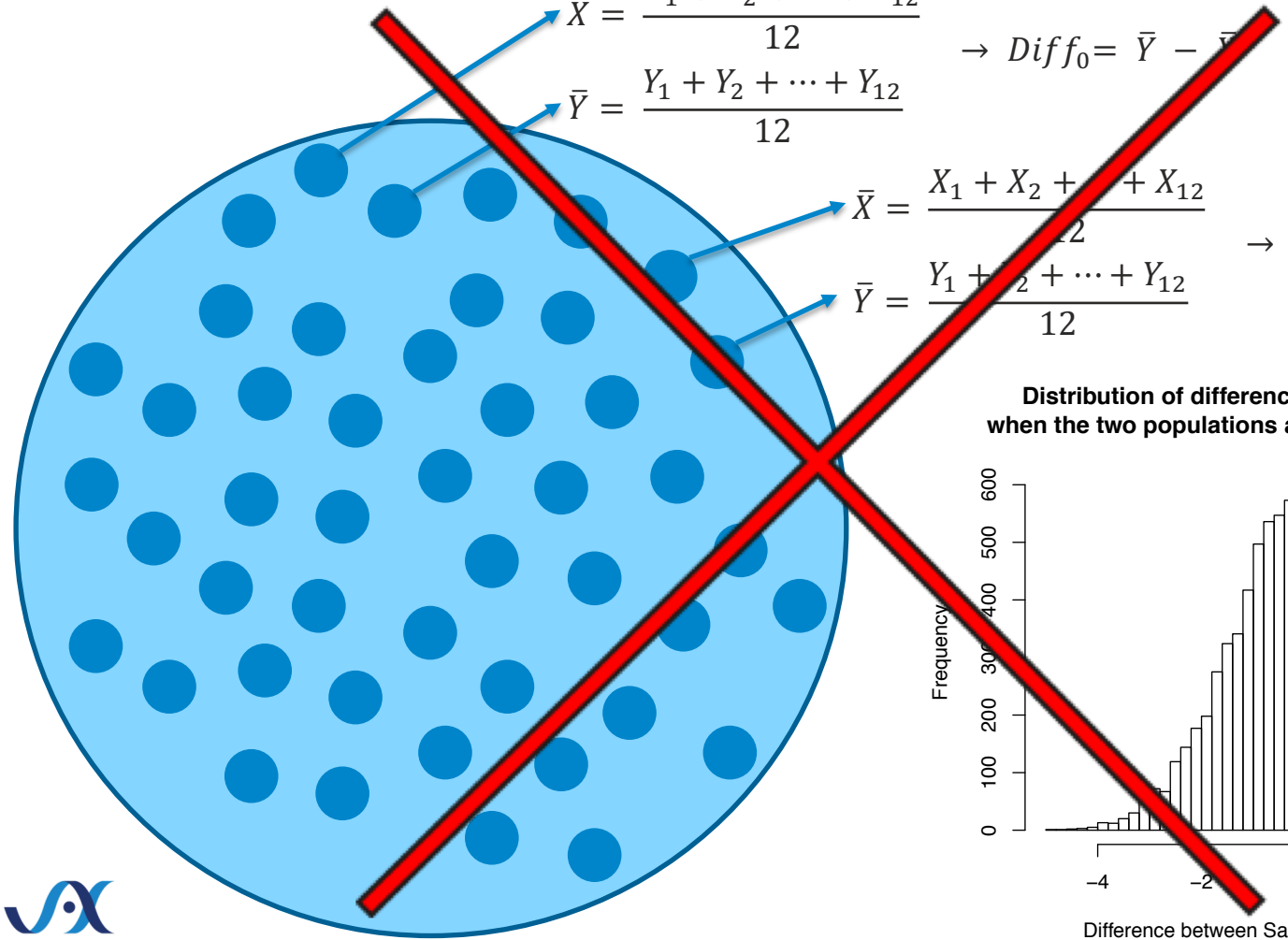
Exercise Break

```
library(tidyverse)

# Load the file that contains the bodyweight data for 12 high fat diet mice and
# 12 control mice. These are our two samples.
SAMPLES <- read_csv('https://raw.githubusercontent.com/genomicsclass/dagdata/master/inst/extdata/femaleMiceWeights.csv')

# Load the population data, remember that in practice we do not have access to
# the population we're interested in. This dataset has both male and female
# data. We will subset it because we are interested only in the female data.
POPULATIONS <- read_csv('https://raw.githubusercontent.com/genomicsclass/dagdata/master/inst/extdata/mice_pheno.csv')
FEMALE_MICE <- POPULATIONS %>%
  filter(Sex == 'F')

CONTROL_POPULATION <- FEMALE_MICE$Bodyweight[FEMALE_MICE$Diet == 'chow']
HIGHFAT_POPULATION <- FEMALE_MICE$Bodyweight[FEMALE_MICE$Diet == 'hf']

CONTROL_SAMPLE <- SAMPLES$Bodyweight[SAMPLES$Diet == 'chow']
HIGHFAT_SAMPLE <- SAMPLES$Bodyweight[SAMPLES$Diet == 'hf']

# 1) What is the control sample mean, standard deviation, and variance?

# 2) What is the high fat diet sample mean, standard deviation, and variance?

# 3) What is the control population mean, standard deviation, and variance?

# 4) What is the high fat diet population mean and standard deviation?

# 5) Is the mean of the high fat diet population different from the control population?
```

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_{12}}{12}$$

$$\bar{Y} = \frac{Y_1 + Y_2 + \cdots + Y_{12}}{12}$$

$$\rightarrow Diff_0 = \bar{Y} - \bar{X}$$

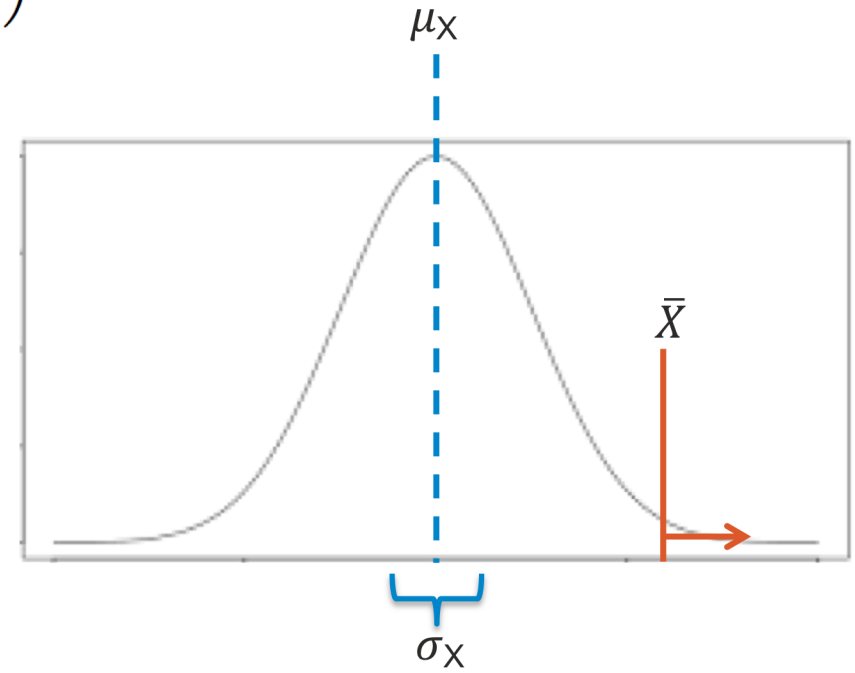


Distribution of differences between sample means when the two populations are identical (null distribution)

Normal Distribution

$$\Pr(a < x < b) = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) dx$$

- If a random variable is normally distributed we can calculate the exact probability of observing such an extreme value
- How do we know if our statistic/estimator/RV is normally distributed?

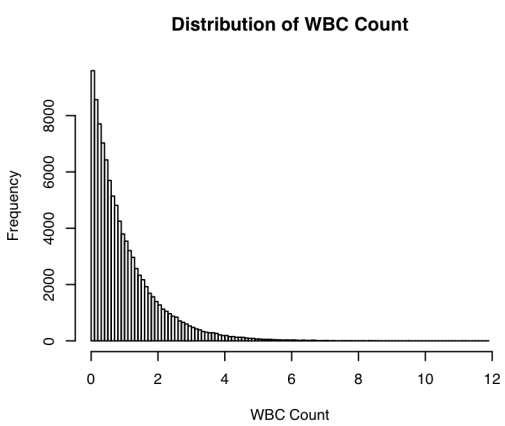


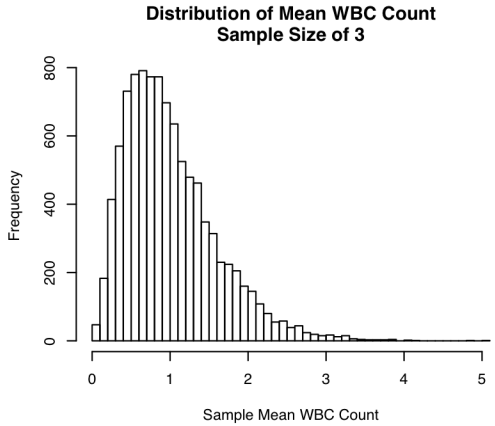
Central Limit Theorem (CLT)

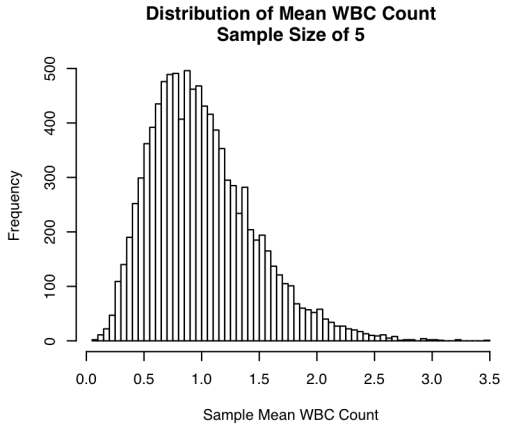
- If your sample size is large enough, the distribution of the sample mean will be normal, even if the distribution of the population is not.
- Y may not be normally distributed, but if N is large enough $\bar{Y}$ will be.

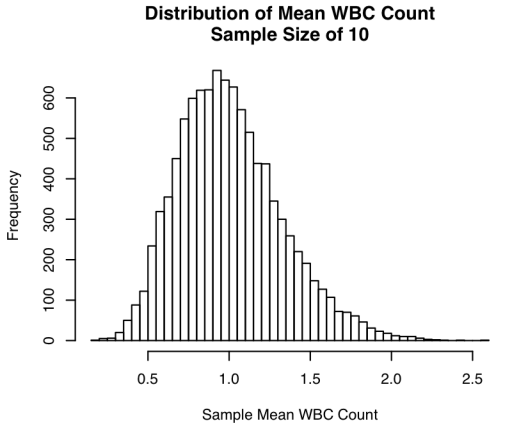
Really?

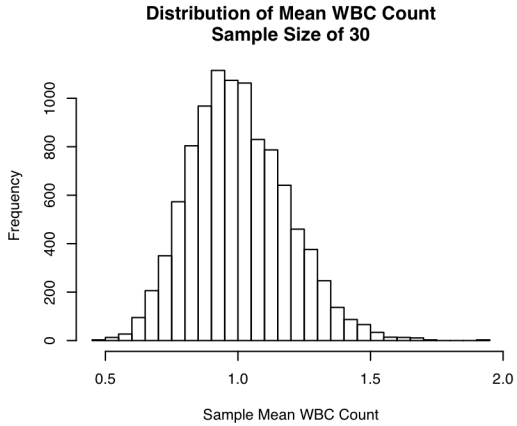
```
set.seed(1)
WBC_count <- rexp(100000, rate = 1)
hist(WBC_count, breaks = 100, main = 'Distribution of WBC Count', xlab = 'WBC Count')

REPS <- 10000

set.seed(1)
SIZE <- 3 # See what happens as you increase this number
Means <- vector(mode = 'numeric', length = REPS)
for(i in 1:REPS){
  Means[i] <- mean(sample(WBC_count, SIZE))
}
hist(
  Means,
  breaks = 50,
  main = paste0('Distribution of Mean WBC Count\nSample Size of ', SIZE),
  xlab = 'Sample Mean WBC Count'
)
```

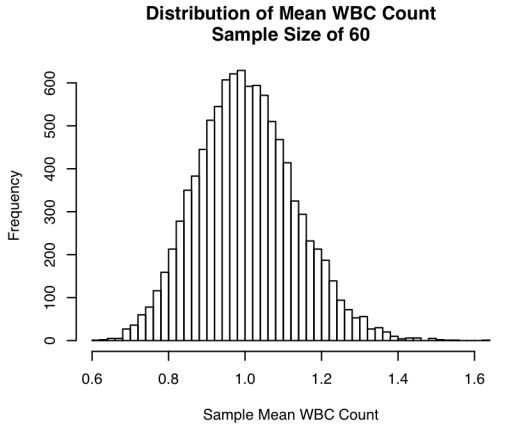

Distribution of WBC Count
Frequency
0
2000
4000
6000
8000
0
2
4
6
8
10
12
WBC Count


Distribution of Mean WBC Count
Sample Size of 3
Frequency
0
200
400
600
800
0
1
2
3
4
5
Sample Mean WBC Count


Distribution of Mean WBC Count
Sample Size of 5
Frequency
0
100
200
300
400
500
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Sample Mean WBC Count


Distribution of Mean WBC Count
Sample Size of 10
Frequency
0
100
200
300
400
500
600
0.5
1.0
1.5
2.0
2.5
Sample Mean WBC Count


Distribution of Mean WBC Count
Sample Size of 30
Frequency
0
200
400
600
800
1000
0.5
1.0
1.5
2.0
Sample Mean WBC Count


Distribution of Mean WBC Count
Sample Size of 60
Frequency
0
100
200
300
400
500
600
0.6
0.8
1.0
1.2
1.4
1.6
Sample Mean WBC Count

More Notes on CLT

- If N is large: $\bar{Y} \sim N(\mu_Y, {}^{\sigma_Y}/_{\sqrt{N}})$
  - Standard Error of $\bar{Y}$: $s_{\bar{Y}} = \frac{\sigma_Y}{\sqrt{N}} \approx \frac{s_Y}{\sqrt{N}}$
    - The standard deviation of the distribution of a random variable is often called the standard error
  - Mean of $\bar{Y}$: μ_Y
- Note that, the larger the sample size, the lower the variation in $\bar{Y}$.

Some Properties of Random Variables

- If you add a constant to a random variable its mean shifts by that constant and its variance/standard deviation remain unchanged.
  - If, $X \sim N(\mu_X, \sigma_X)$
  - Then, $X - a \sim N(\mu_X - a, \sigma_X)$
- If you multiply a constant by a random variable, the mean and standard deviation are both scaled by that constant
  - If, $X \sim N(\mu_X, \sigma_X)$
  - Then, $a * X \sim N(a * \mu_X, |a| * \sigma_X)$
    - Note that standard deviations and other measures of variation are always positive
- These two rules allow us to convert our sample mean to what is often called a Z-score.

Making the CLT more Useful:
Standard Normal Distribution and Z-Scores

$$\bar{Y} \sim N(\mu_Y, {}^{\sigma_Y}/_{\sqrt{N}}) \quad \rightarrow \quad \frac{\bar{Y} - \mu_Y}{\sigma_Y/\sqrt{N}} \sim N(0,1)$$

- If we subtract the assumed population mean (a constant) from our sample mean and divide by the standard deviation, our transformed random variable will have a normal distribution with a mean of zero and a standard deviation of one.
- This is called the standard normal distribution (sometimes a Z-distribution)
- This new variable is often called a Z-score

Exercise Break

```
# 1) Compare the control sample standard deviation and the control
population standard deviation.

# 2) What is the control sample Z-score using the population standard
deviation?

# 3) What about when we use the sample standard deviation?

# 4) For each of these test statistics (answers to #3 and #4 above),
use pnorm() to find the probability of observing such an extreme value
(the p-value).
```

Some Properties of Normal Random Variables

- If you take the sum or difference of two independent normal random variables, the resulting variable is also normally distributed
  - If, $Y \sim N(\mu_Y, \sigma_Y)$ and $X \sim N(\mu_X, \sigma_X)$
  - Then,

$$Y - X \sim N\left(\mu_Y - \mu_X, \sqrt{\sigma_Y^2 + \sigma_X^2}\right)$$

and

$$Y + X \sim N\left(\mu_Y + \mu_X, \sqrt{\sigma_Y^2 + \sigma_X^2}\right)$$

  - The mean is the sum/difference of the two means
  - The variance is *always* the sum of the two variances

Making the CLT more Useful

$$\frac{\bar{Y} - \bar{X}}{\sqrt{\frac{\sigma_Y^2}{N} + \frac{\sigma_X^2}{M}}} \sim N(\mu_Y - \mu_X, 1) \rightarrow \frac{\bar{Y} - \bar{X}}{\sqrt{\frac{\sigma_Y^2}{N} + \frac{\sigma_X^2}{M}}} \sim N(0, 1)$$

- If $\bar{Y}$ and $\bar{X}$ are independent normally distributed random variables, then their difference ($\bar{Y}$ - $\bar{X}$) is also a normally distributed random variable.

- If $\bar{Y}$ and $\bar{X}$ are independent normally distributed random variables, then the variance of their sum (or difference) is the sum of their variances.

Making the CLT more Useful

$$\frac{\bar{Y} - \bar{X}}{\sqrt{\frac{\sigma_Y^2}{N} + \frac{\sigma_X^2}{M}}} \sim N(0,1) \rightarrow \frac{\bar{Y} - \bar{X}}{\sqrt{\frac{s_Y^2}{N} + \frac{s_X^2}{M}}} \overset{*}{\sim} N(0,1)$$

- We estimate the population variances/standard deviations with the sample statistics

* Since we are estimating the population variance/standard deviation, this is actually not normally distributed, but *approximately* normally distributed. For this reason, we generally use a different probability distribution, called the t-distribution, to calculate confidence intervals and p-values. The t-distribution is similar to the standard normal distribution but it varies by sample size. At large samples it is very close to the standard normal distribution.

Useful note

If the sizes of the two samples are equal (N = M) then:

$$\frac{\bar{Y} - \bar{X}}{\sqrt{\frac{s_Y^2}{N} + \frac{s_X^2}{M}}} = \sqrt{N} * \frac{\bar{Y} - \bar{X}}{\sqrt{s_Y^2 + s_X^2}}$$

$$\frac{\bar{Y} - \bar{X}}{\sqrt{\frac{\sigma_Y^2}{N} + \frac{\sigma_X^2}{M}}} = \sqrt{N} * \frac{\bar{Y} - \bar{X}}{\sqrt{\sigma_Y^2 + \sigma_X^2}}$$

Exercise Break

1) What is the Z-score of the difference between the high fat diet sample mean and the control diet sample mean using the population standard deviations?

2) What is this value if the sample standard deviations are used instead?

3) Use pnorm() to find the p-values for the answers to #3 and #4. Are they the same?

CLT Summary

- When to use:
  - Large sample size, and working with means (or differences and sums of means)
- One sample: $\frac{\bar{Y} - \mu_Y}{s_Y/\sqrt{N}} \overset{*}{\sim} N(0,1)$
  - The true population mean is unknown, μ_Y is what it is assumed to be under the null hypothesis
- Difference of two groups: $\frac{\bar{Y} - \bar{X}}{\sqrt{\frac{s_Y^2}{N} + \frac{s_X^2}{M}}} \overset{*}{\sim} N(0,1)$
  - This is under the null hypothesis of no difference ($\mu_Y - \mu_X = 0$)

Lies, lies, lies!

$$\frac{\bar{Y} - \mu_Y}{\sigma_Y/\sqrt{N}} \sim N(0,1)$$

$$\frac{\bar{Y} - \mu_Y}{s_Y/\sqrt{N}} \overset{*}{\sim} N(0,1)$$

$$\frac{\bar{Y} - \bar{X}}{\sqrt{\frac{\sigma_Y^2}{N} + \frac{\sigma_X^2}{M}}} \sim N(0,1)$$

$$\frac{\bar{Y} - \bar{X}}{\sqrt{\frac{s_Y^2}{N} + \frac{s_X^2}{M}}} \overset{*}{\sim} N(0,1)$$

Student's t-Distribution

- Ratios of random variables are not normally distributed
  - They are more likely to take extreme values so the tails of the distribution are larger
- The t-distribution takes is defined by a parameter called "degrees of freedom" which is closely related to sample size
- The t-distribution looks a lot like the standard normal distribution and at large sample sizes they are almost identical
- What this means in practice:
  - We use t-statistics rather than Z-scores
  - We use `pt` (the "t" is for t-distribution) rather than `pnorm`

Exercise Break

Exercises #1 and #2 are repeats, if you still have the values in your R working environment you don't have to bother re-calculating them

1) Again, find the test statistic for the difference between the high fat and control means using the sample standard deviation. This is the test statistic we would be using in practice since we don't have access to population parameters.

2) If we assume this test statistic has a standard normal distribution, what is its p-value? Use the pnorm() function.

3) If we assume this test statistic has a standard normal distribution, what is its p-value? Use the pt() function with degrees of freedom of 22 (df = 22):

```
2*(1 – pt(abs(TEST_STATISTIC), df = 22))
```

4) Are the values different? Why?

The t-Distribution?

- Looks like the standard normal distribution, but has heavier tails (increased probability of extreme values)
- Used when sample size is small but the distribution of Y (**not** $\bar{Y}$) can be assumed to be approximately normal
- Takes just one parameter called the “degrees of freedom” which is closely linked to sample size
- At large sample sizes, the t-distribution approaches the standard normal distribution