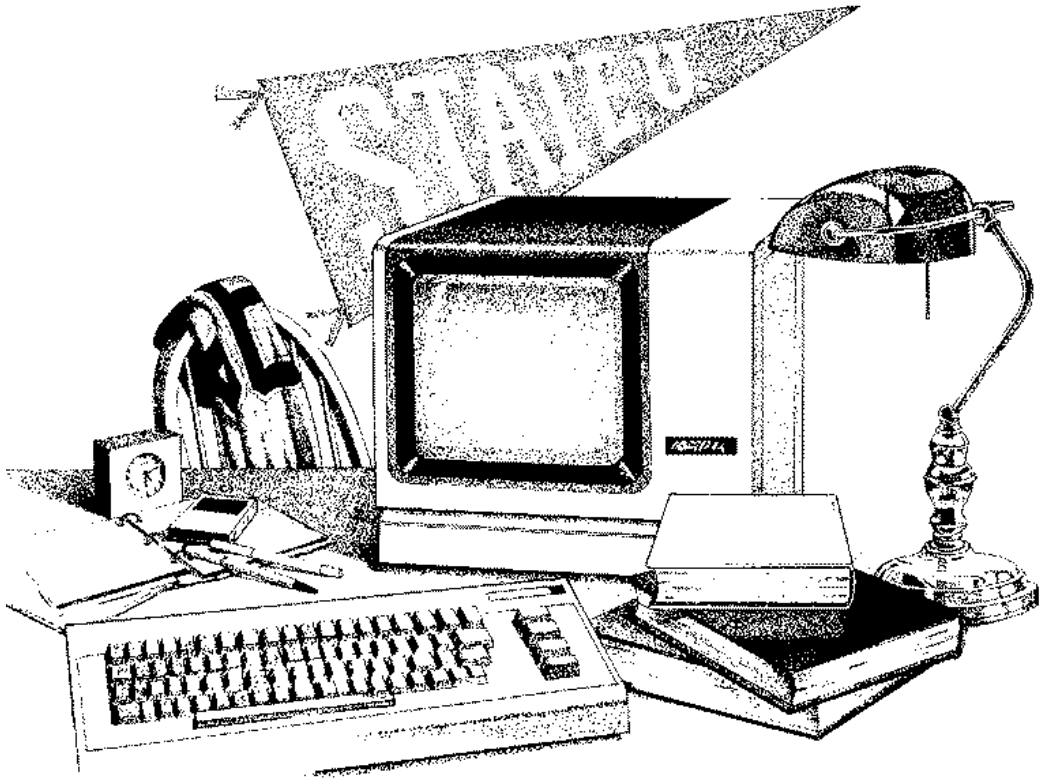
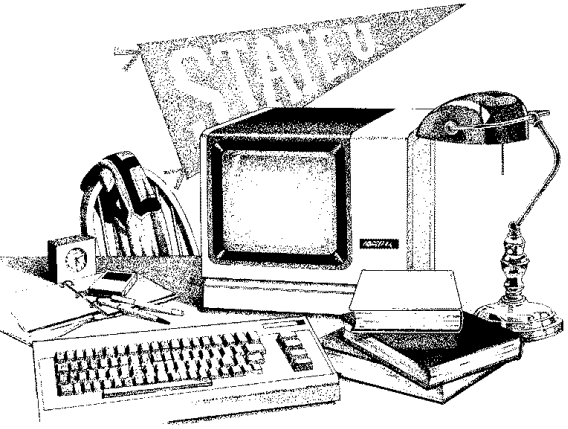
# AP Statistics Chapter 23

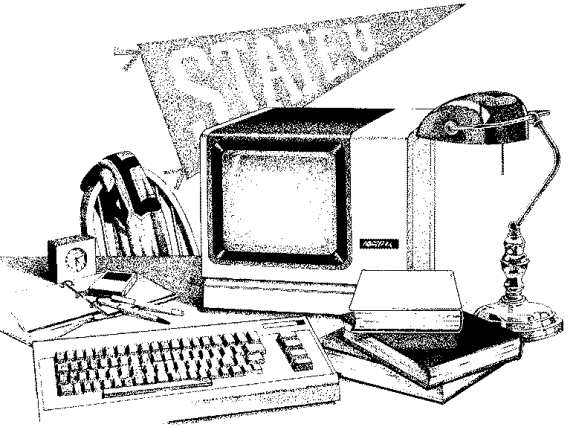
Comparing Means





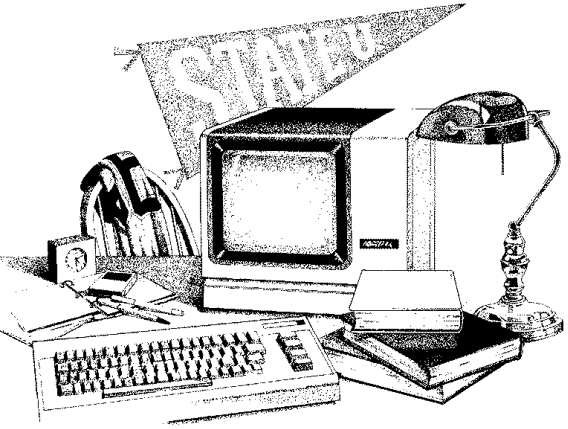

# Objectives:

- Two-sample t methods
- Two-Sample t Interval for the Difference Between Means
- Two-Sample t Test for the Difference Between Means
- Pooling
- Pooling-t Methods




# Two-Sample Problems

- The goal of inference is to compare the responses to two treatments or to compare the characteristics of two populations.
- Have a separate sample from each treatment or each population.
- The responses of each group are independent of those in the other group.

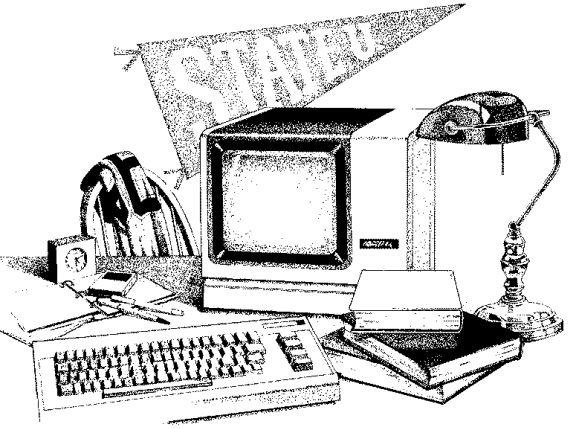



# Definitions

## Two Samples: Independent

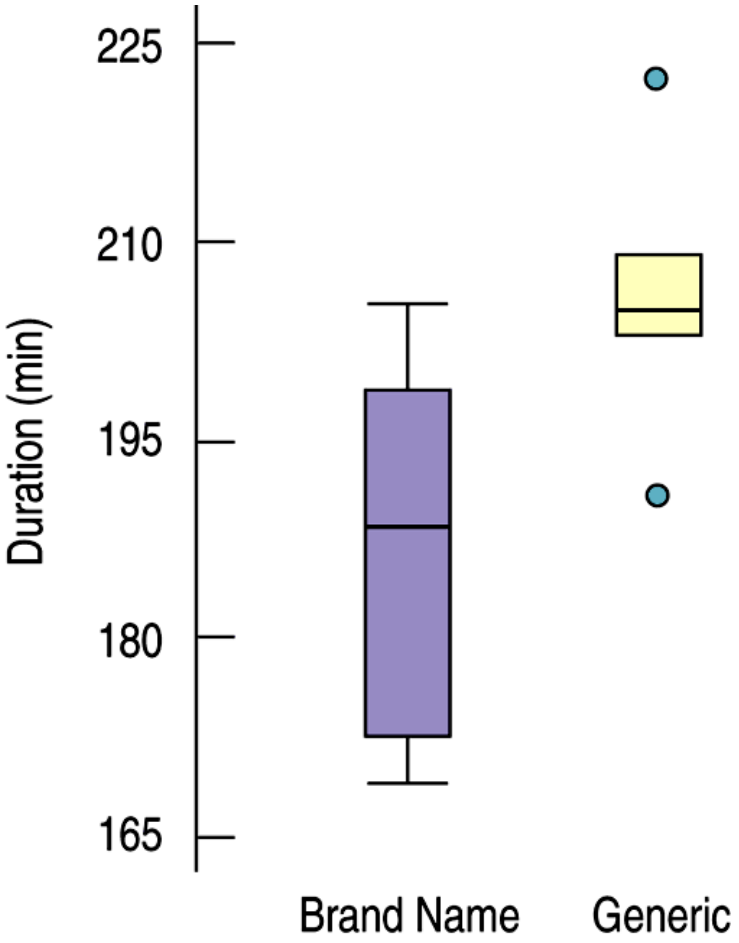
The sample values selected from one population are not related or somehow paired with the sample values selected from the other population.

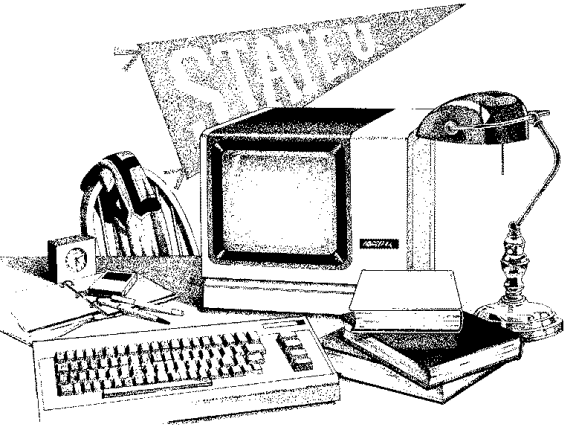
If the values in one sample are related to the values in the other sample, the samples are dependent. Such samples are often referred to as matched pairs or paired samples.




# Plot the Data

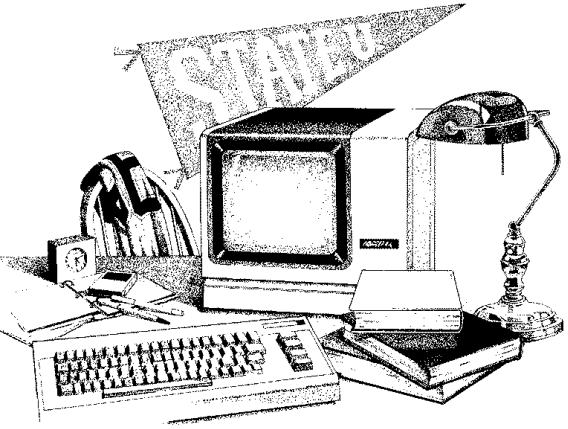
- The natural display for comparing two groups is boxplots of the data for the two groups, placed side-by-side. For example:






# Comparing Two Means

- Once we have examined the side-by-side boxplots, we can turn to the comparison of two means.
- Comparing two means is not very different from comparing two proportions.
- This time the parameter of interest is the difference between the two means, $\mu_1 - \mu_2$.



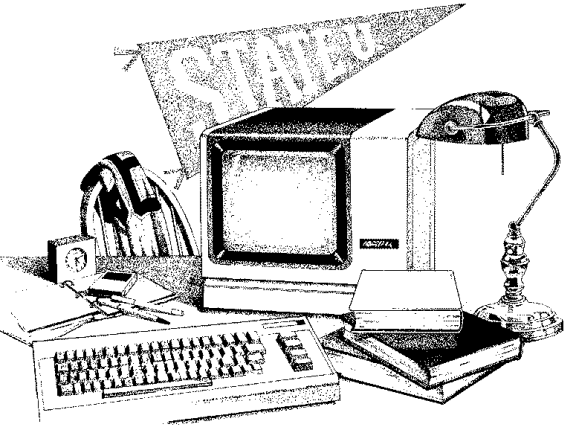

# Comparing Two Means (cont.)

- Remember that, for independent random quantities, variances add.
- So, the standard deviation of the difference between two sample means is

$$SD(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

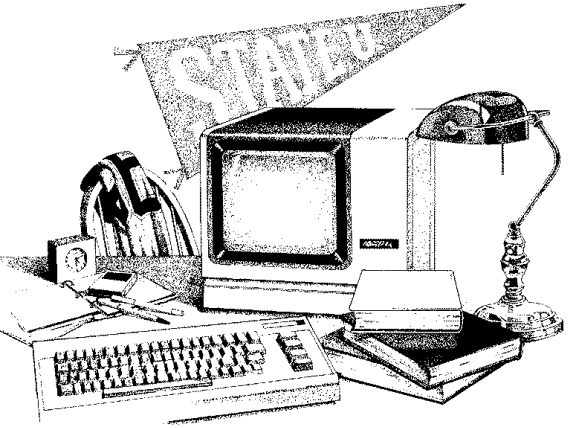
- We still don't know the true standard deviations of the two groups, so we need to estimate and use the standard error

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$




# Comparing Two Means (cont.)

- Because we are working with means and estimating the standard error of their difference using the data, we shouldn't be surprised that the sampling model is a Student's *t*.
  - The confidence interval we build is called a two-sample *t*-interval (for the difference in means).
  - The corresponding hypothesis test is called a two-sample *t*-test.




# Sampling Distribution for the Difference Between Two Means

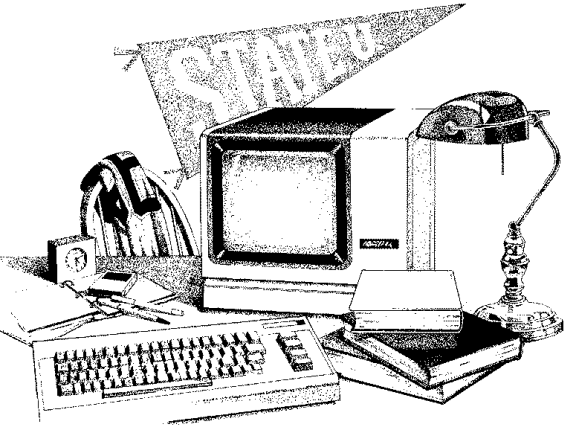
- When the conditions are met, the standardized sample difference between the means of two independent groups

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{SE(\bar{y}_1 - \bar{y}_2)}$$

can be modeled by a Student's *t*-model with a number of degrees of freedom found with a special formula.

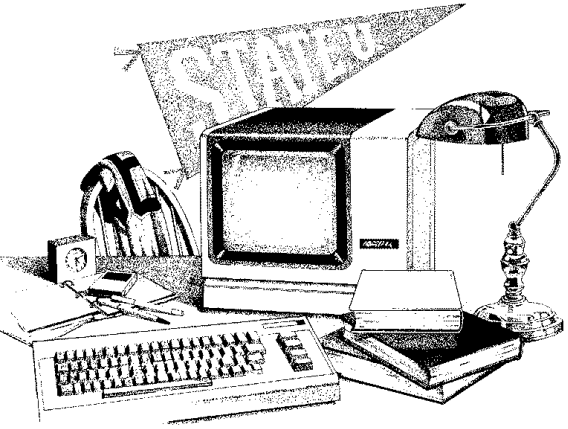
- We estimate the standard error with

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$



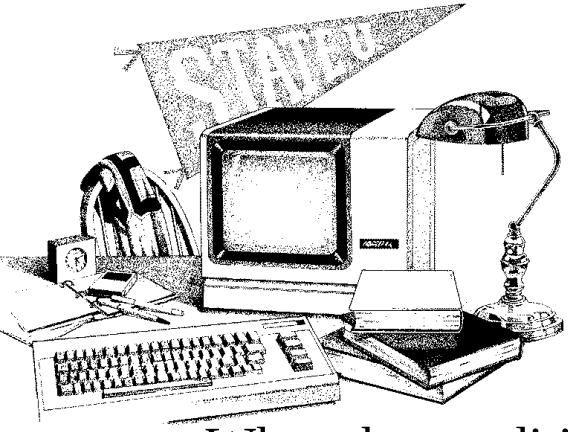

# Assumptions and Conditions

- Independence Assumption (Each condition needs to be checked for both groups.):
  - Randomization Condition: Were the data collected with suitable randomization (representative random samples or a randomized experiment)?
  - 10% Condition: We don't usually check this condition for differences of means. We will check it for means only if we have a very small population or an extremely large sample.



# Assumptions and Conditions (cont.)

- Normal Population Assumption:
  - Nearly Normal Condition: This must be checked for *both* groups. A violation by either one violates the condition. Both samples come from a normal population, or samples are large (>40), or samples are medium (15-40) and plots show little skewness and no outliers, or samples are small (<15) and plots show no skewness and no outliers.
- Independent Groups Assumption: The two groups we are comparing must be independent of each other. (See Chapter 25 if the groups are not independent of one another…)




# Two-Sample *t*-Interval

When the conditions are met, we are ready to find the confidence interval for the difference between means of two independent groups, $\mu_1 - \mu_2$.

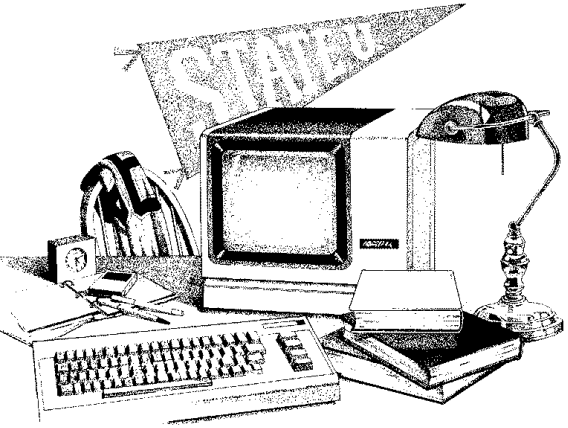
The confidence interval is

$$\left(\bar{y}_1 - \bar{y}_2\right) \pm t^*_{df} \times SE\left(\bar{y}_1 - \bar{y}_2\right)$$

where the standard error of the difference of the means is

$$SE\left(\bar{y}_1 - \bar{y}_2\right) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

The critical value $t^*_{df}$ depends on the particular confidence level, $C$, that you specify and on the number of degrees of freedom, which we get from the sample sizes and a special formula.

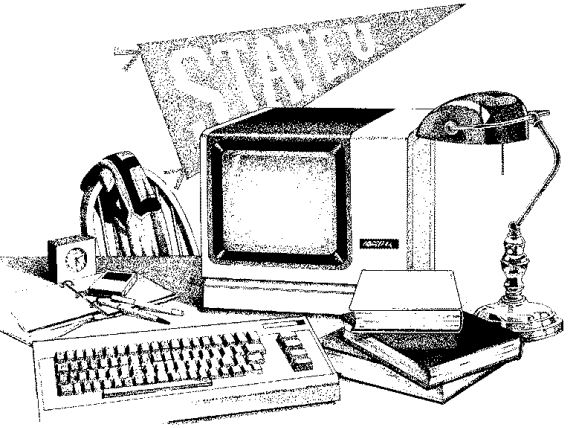



# Degrees of Freedom

- The special formula for the degrees of freedom for our *t* critical value is a bear:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

- Because of this, we will use this estimate: *df* = smaller of $n_1 - 1$ and $n_2 - 1$.

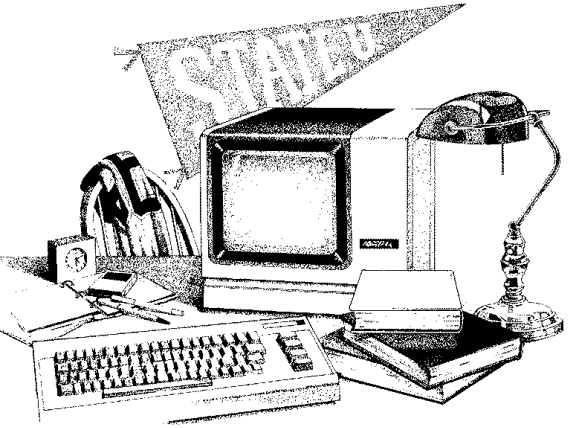



# Two-Sample t Procedures

- **Degrees of freedom: Use this estimate: *df* = smaller of $n_1 - 1$ and $n_2 - 1$.**

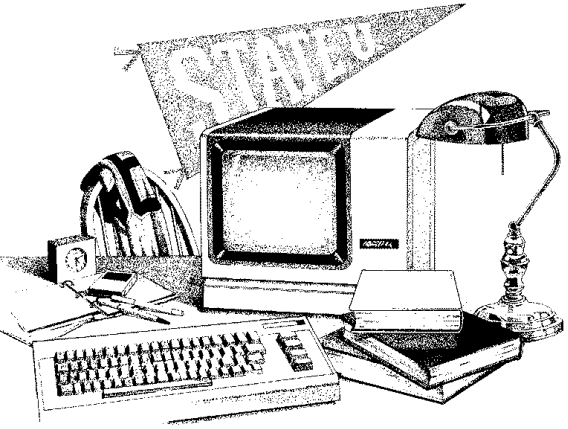
- **Confidence interval for $\mu_1$-$\mu_2$:**

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$



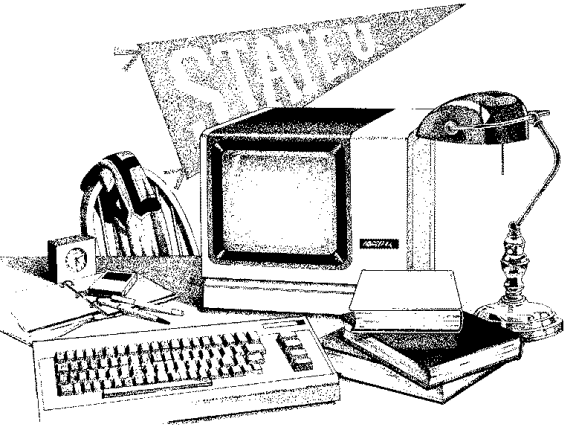
# Example: Two-Sample t-Interval for Means

- Manufacturers make claims about their products and usually try to convince you that their product is better than that of a competitor. Most brands of paper towels claim to pick up more liquid than any other brand. How much of a difference, on average, can be expected between Brand A and Brand B? This calls for a confidence interval to find the true difference, $\mu_A$-$\mu_B$, between the mean number of milliliters of water absorbed by each towel. A random sample of 16 of each type of towel was tested for absorbency. The mean number of ml. for Brand A was 15.625 ml. with a standard deviation of 3.12 ml. while for Brand B the mean was 14 ml. with a standard deviation of 2.53 ml. Find 95% confidence interval for $\mu_A$-$\mu_B$.



# Solution

1. Check conditions
   - Independence: Two different brands, therefore independent.
   - Randomization: Data was from random samples.
   - 10% Condition: There are more than 160 towels of each brand.
   - Normality: Sample size medium (15-40), assume histogram of samples show little skewness and no outliers.




# Solution

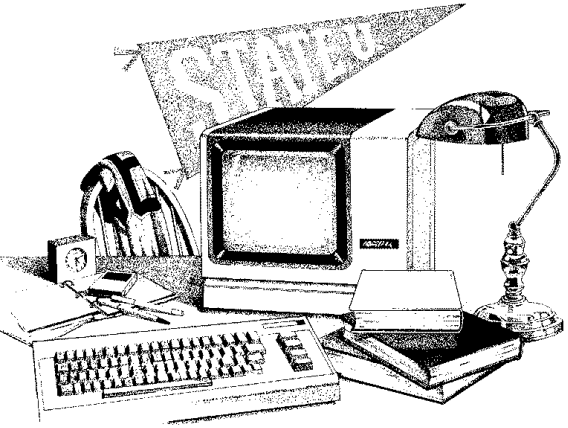
2. Calculate 95% Confidence Interval

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$95\% \quad t^*_{15} = 2.13$$

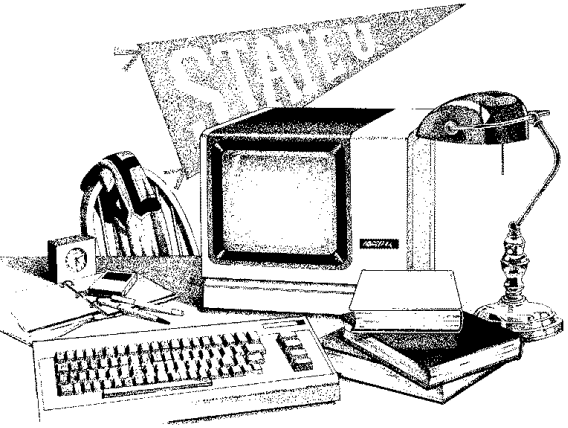
$$(\bar{x}_1 - \bar{x}_2) \pm t^* \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = (15.625 - 14) \pm (2.13)\sqrt{\frac{3.12^2}{16} + \frac{2.53^2}{16}} = 1.625 \pm 2.139$$

$$(-.514, 3.764)$$



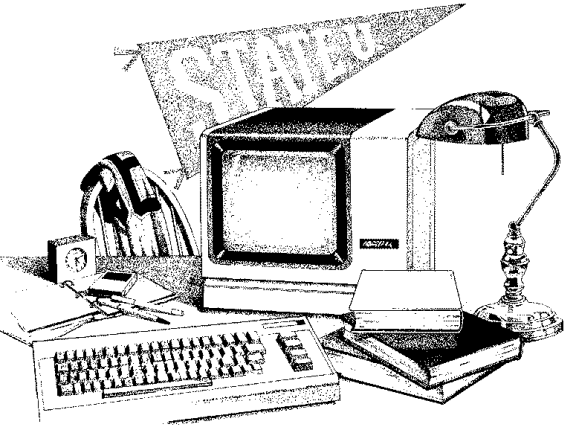
# Solution

- Conclusion
  - We are 95% confident that the mean difference in the number of milliliters of water absorbed by each towel is -.514 mL and 3.764 mL.



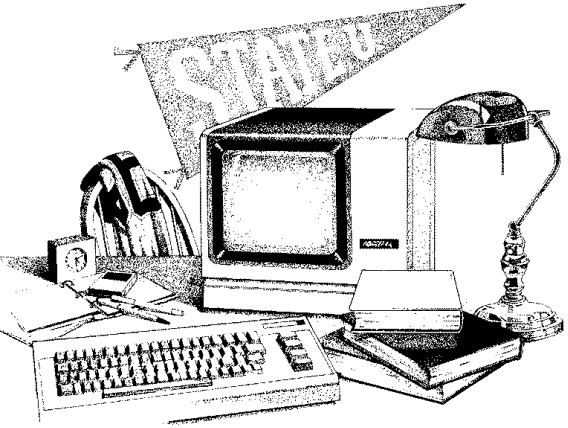

# TI-84 Solution

- (-.4296, 3.6796)
- df = 28.772




# Testing the Difference Between Two Means

- The hypothesis test we use is the two-sample *t*-test for means.
- The conditions for the two-sample *t*-test for the difference between the means of two independent groups are the same as for the two-sample *t*-interval.



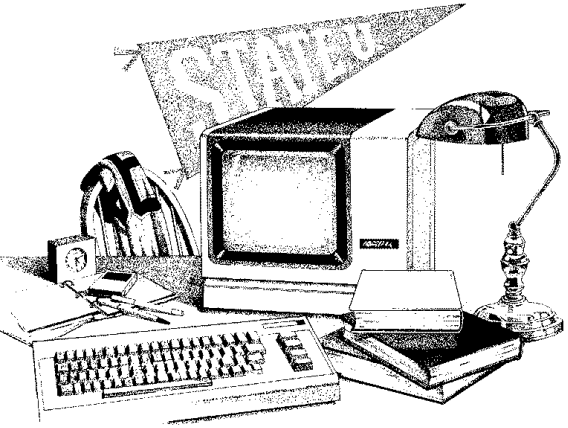

# A Test for the Difference Between Two Means

- We test the hypothesis $H_0: \mu_1 - \mu_2 = \Delta_0$, where the hypothesized difference, $\Delta_0$, is almost always 0, using the statistic

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - \Delta_0}{SE(\bar{y}_1 - \bar{y}_2)}$$

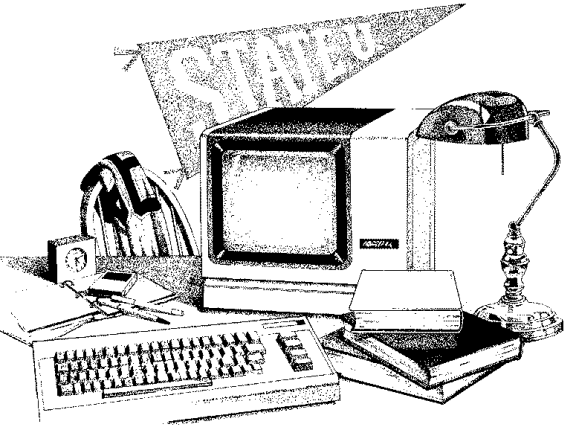
- The standard error is $SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

- When the conditions are met and the null hypothesis is true, this statistic can be closely modeled by a Student's *t*-model with a number of degrees of freedom given by the estimate: *df* = smaller of $n_1 - 1$ and $n_2 - 1$. We use that model to obtain a P-value.




# Assumptions/Conditions

1. The two samples are *independent*.
2. Both samples are *simple random samples*.
3. *Population Size, 10% condition*
4. *Normality*. Both samples come from a normal population, or samples are large (>40), or samples are medium (15-40) and plots show little skewness and no outliers, or samples are small (<15) and plots show no skewness and no outliers.

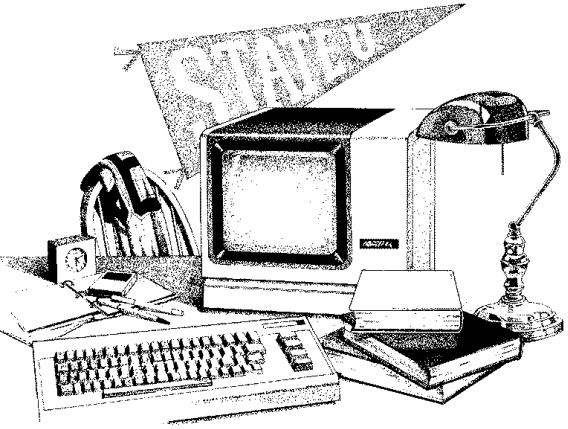


# Hypotheses

- $H_0$: $\mu_1=\mu_2$
- $H_a$: $\mu_1\neq\mu_2$, or $\mu_1>\mu_2$, or $\mu_1<\mu_2$

**Or, equivalently**

- $H_0$: $\mu_1-\mu_2=0$
- $H_a$: $\mu_1-\mu_2\neq0$, or $\mu_1-\mu_2>0$, or $\mu_1-\mu_2<0$

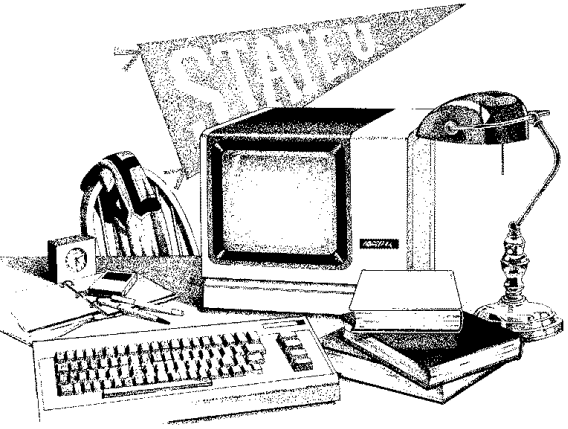



# Two-Sample t Procedures

- Degrees of freedom: Use this estimate: *df* = smaller of $n_1 - 1$ and $n_2 - 1$.

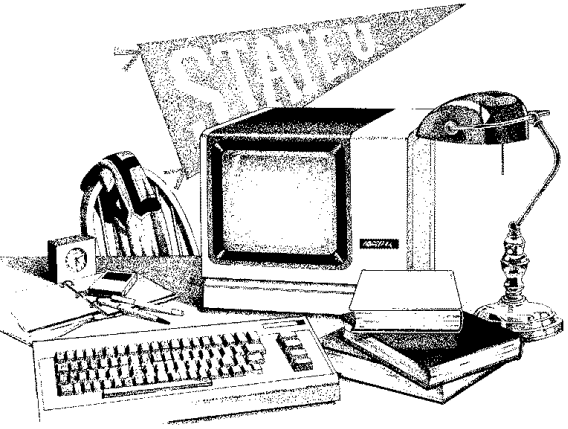
- Two-sample t statistic for $H_0$: $\mu_1=\mu_2$:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$



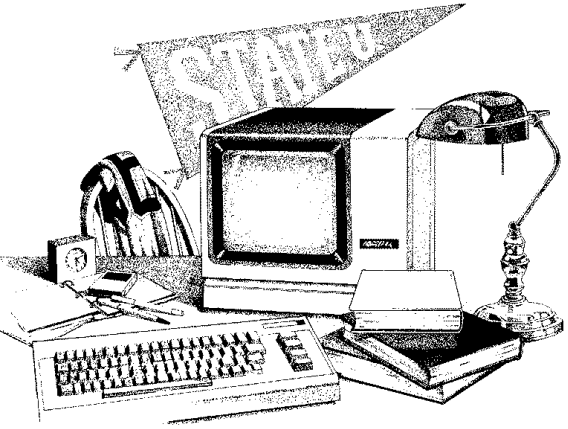

# Example: Two-Sample t-Test for Means

- It is a common belief that women tend to live longer than men. Random samples from the death records for men and women in Montgomery County were taken and age at the time of death recorded. The average age of the 48 males was 68.33 years with a standard deviation of 12.49 years, while the average age of the 40 females was 78.7 years with a standard deviation of 16.43 years. Do women in this country tend to live longer than men?



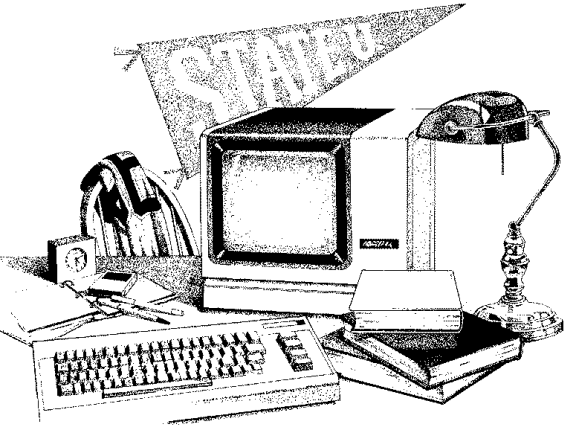

# Solution

1. Check Conditions
   - Independence: Males and females are independent groups.
   - Randomization: Random samples were taken.
   - 10% Condition: There are more than 480 and 400 death records.
   - Normality: Both samples are large ($n \geq 40$).



# Solution

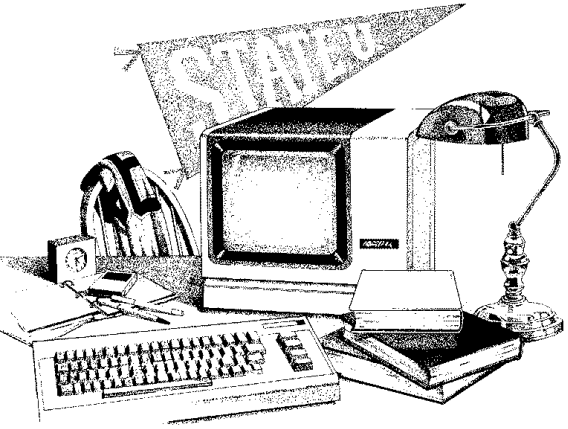
2. Hypothesis
   - Null Hypothesis: $H_0: \mu_M = \mu_F$ (There is no difference in the mean ages at death between men and women in this country).
   - Alternative Hypothesis: $H_a: \mu_M < \mu_F$ ( The mean age at death for men in this country is less than that for women).



# Solution

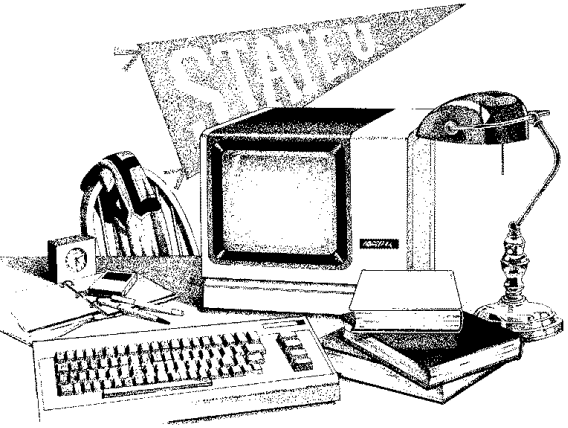
3. Calculate the Test Statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(68.33 - 78.7) - 0}{\sqrt{\frac{12.49^2}{48} + \frac{16.43^2}{40}}} = \frac{-10.37}{3.162} = -3.28$$



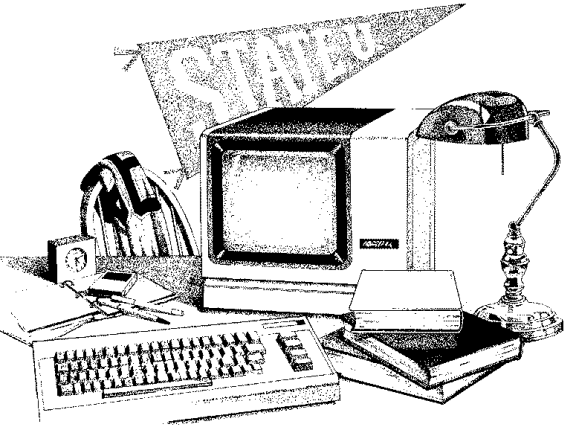

# Solution

- Calculate P-value
  - P-value = $P(t_{39} < -3.28) = .0011$
- Conclusion
  - With a P-value of .0011, very small, we reject the null hypothesis and conclude there is sufficient evidence that women live longer than men.



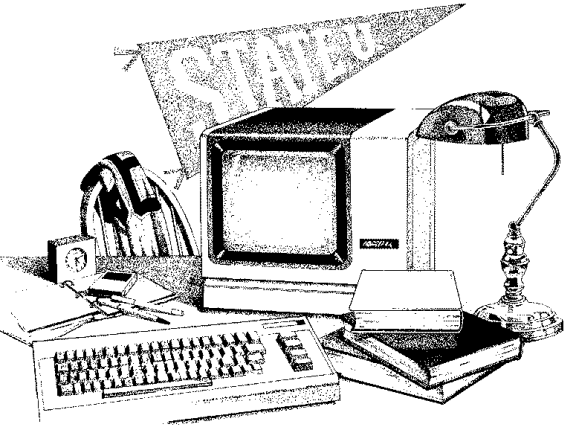

# TI-84 Solution

- t = -3.28
- P-value = .000803
- df = 71.8



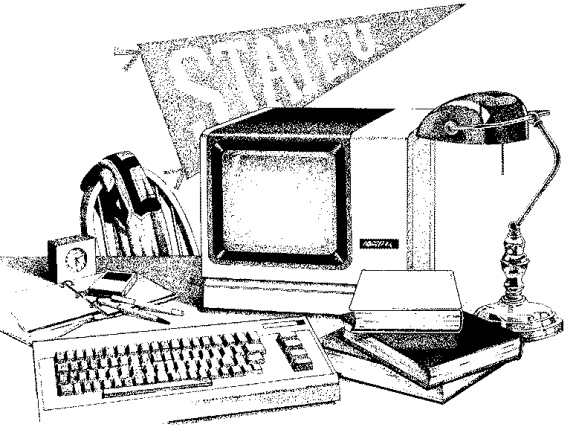

# Back Into the Pool

- Remember that when we know a proportion, we know its standard deviation.
  - Thus, when testing the null hypothesis that two proportions were equal, we could assume their variances were equal as well.
  - This led us to pool our data for the hypothesis test.



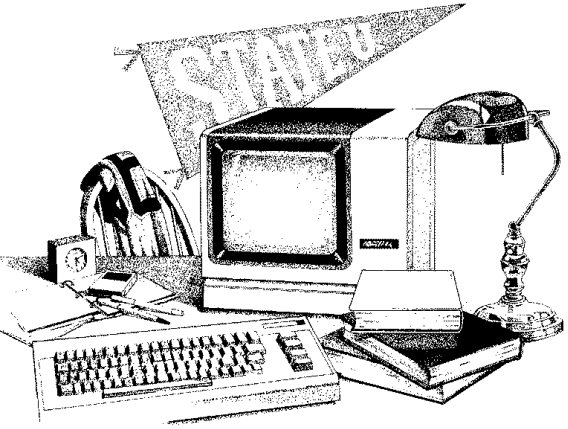

# Back Into the Pool (cont.)

- For means, there is also a pooled *t*-test.
  - Like the two-proportions *z*-test, this test assumes that the variances in the two groups are equal.
  - But, be careful, there is no link between a mean and its standard deviation…




# Back Into the Pool (cont.)

- If we are willing to *assume* that the variances of two means are equal, we can pool the data from two groups to estimate the common variance and make the degrees of freedom formula much simpler.
- We are still estimating the pooled standard deviation from the data, so we use Student's *t*-model, and the test is called a pooled *t*-test (for the difference between means).




# *The Pooled *t*-Test

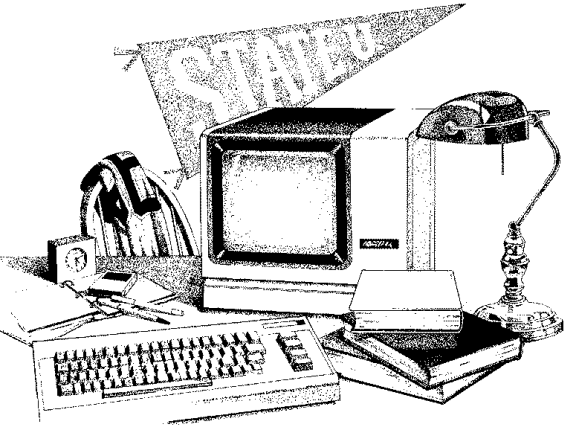
- If we assume that the variances are equal, we can estimate the common variance from the numbers we already have:

$$s^2_{pooled} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

- Substituting into our standard error formula, we get:

$$SE_{pooled}(\bar{y}_1 - \bar{y}_2) = s_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- Our degrees of freedom are now $df = n_1 + n_2 - 2$.

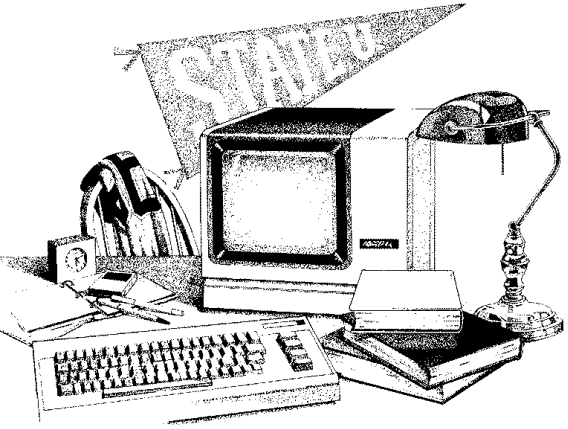



# *The Pooled *t*-Test and Confidence Interval for Means

- The conditions for the pooled *t*-test and corresponding confidence interval are the same as for our earlier two-sample *t* procedures, with the additional assumption that the variances of the two groups are the same.
- For the hypothesis test, our test statistic is $t = \dfrac{(\bar{y}_1 - \bar{y}_2) - \Delta_0}{SE_{pooled}(\bar{y}_1 - \bar{y}_2)}$

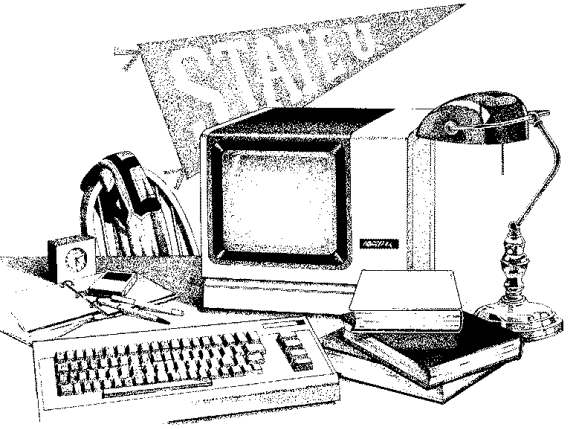
  which has $df = n_1 + n_2 - 2$.
- Our confidence interval is

$$(\bar{y}_1 - \bar{y}_2) \pm t^*_{df} \times SE_{pooled}(\bar{y}_1 - \bar{y}_2)$$



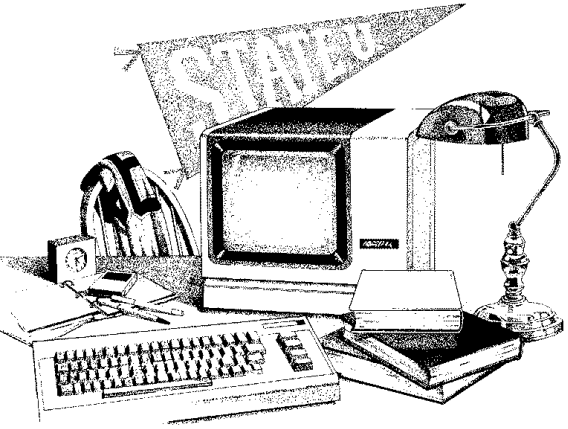

# Is the Pool All Wet?

- So, when *should* you use pooled-*t* methods rather than two-sample *t* methods? Never. (Well, hardly ever.)
- Because the advantages of pooling are small, and you are allowed to pool only rarely (when the equal variance assumption is met), **don't**.
- It's never wrong ***not*** to pool.



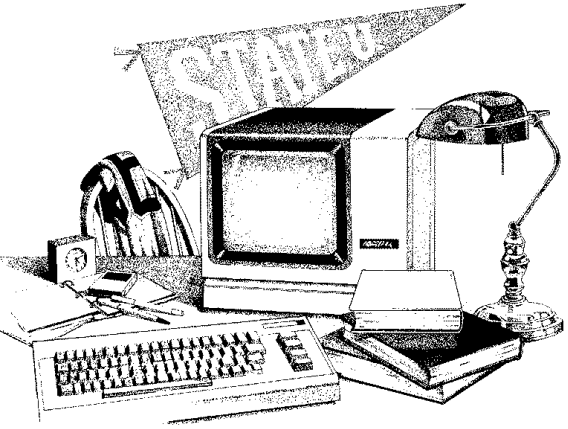

# Why Not Test the Assumption That the Variances Are Equal?

- There is a hypothesis test that would do this.
- But, it is very sensitive to failures of the assumptions and works poorly for small sample sizes—just the situation in which we might care about a difference in the methods.
- So, the test does not work when we would need it to.



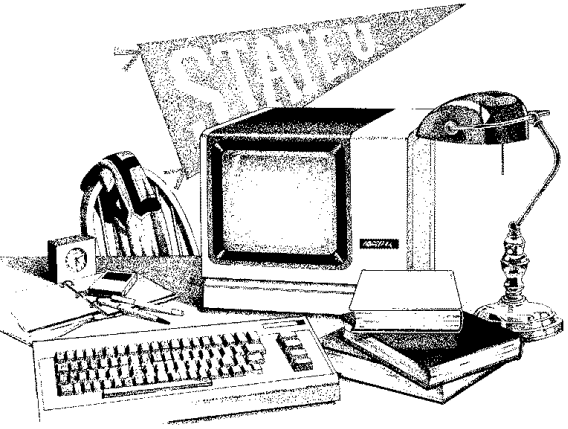

# Is There Ever a Time When Assuming Equal Variances Makes Sense?

- Yes. In a randomized comparative experiment, we start by assigning our experimental units to treatments at random.
- Each treatment group therefore begins with the same population variance.
- In this case assuming the variances are equal is still an assumption, and there are conditions that need to be checked, but at least it’s a plausible assumption.




# What Can Go Wrong?

- Watch out for paired data.
  - The Independent Groups Assumption deserves special attention.
  - If the samples are not independent, you can't use two-sample methods.
- Look at the plots.
  - Check for outliers and non-normal distributions by making and examining boxplots.




# What have we learned?

- We've learned to use statistical inference to compare the means of two independent groups.
  - We use *t*-models for the methods in this chapter.
  - It is still important to check conditions to see if our assumptions are reasonable.
  - The standard error for the difference in sample means depends on believing that our data come from independent groups, but pooling is not the best choice here.
  - Once again, we've see new can add variances.
- The reasoning of statistical inference remains the same; only the mechanics change.