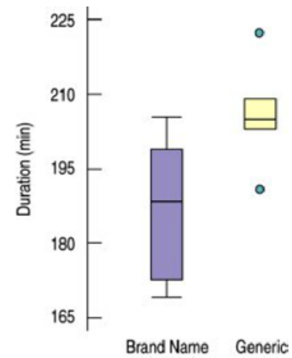
Chapter 24 Summary
Comparing Means

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.
- The reasoning of statistical inference remains the same; only the mechanics change.

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, $\mu_1 - \mu_2$.
- 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 $SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- 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 hypothesis 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.
- Normal Population Assumption:
  - Nearly Normal Condition: This must be checked for *both* groups. A violation by either one violates the condition.
- 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.
- 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 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 let technology calculate degrees of freedom for us!

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.
- We test the hypothesis $H_0: \mu_1 - \mu_2 = \Delta_0$, where the hypothesized difference, Δ_0, is almost always 0, using the statistic

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

  The standard error is $SE\left(\bar{y}_1 - \bar{y}_2\right) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- 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 a special formula. We use that model to obtain a P-value.

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.
- 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…
- 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.

*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

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

  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.

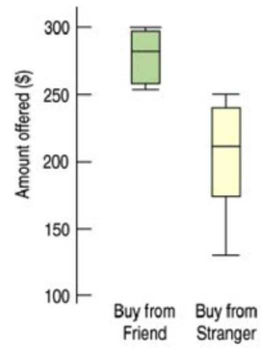
Why Not Test the Assumption That the Variances Are Equal?

- 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.

*Tukey's Quick Test



- To use Tukey's test, one group must have the highest value and the other, the lowest.
- Count how many values in the high group are higher than *all* the values in the lower group.
- Add to this the number of values in the low group that are lower than *all* the values in the higher group. (Count ties as ½.)
- If this total is 7 or more, we can reject the null hypothesis of equal means at $\alpha = 0.05$.
  - The "critical values" of 10 and 13 give us α's of 0.01 and 0.001.
- This is another example of a distribution-free test and a remarkably good test.
- The only assumption it requires is that the two samples be independent.
- Since this test is simple, consider using it as a check of your two-sample *t* results—if the results disagree, check the assumptions.

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.