Comparing Means I

Recap

You tell me -- what is a one-sample *t*-test?

Independent samples *t*-test

- Independent samples *t*-tests are used when you want to compare means from two independent samples
- Samples are independent if observations from one sample do not affect or depend on observations from the other sample. A score from sample 1 shouldn't tell me anything about a score from sample 2 (e.g., whether the score is above or below the sample mean)
- Basically, you want independence of observations within your samples and between your samples for this analysis
- Very common with experimental research (e.g., compare Treatment condition mean with Control condition mean)

Hypotheses

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Alternatively. . . (rearranged formula)

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 - \mu_2 \neq 0$$

We can calculate the test statistic with:

$$t_{df} = \frac{\bar{X}_1 - \bar{X}_2}{SE_{meandifference}}$$

Technically, the formula is:

$$t_{df} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{SE_{meandifference}}$$

- $(\bar{X}_1 - \bar{X}_2)$ is our observed difference in (independent) means
- Plug in values for $(\mu_1 - \mu_2)$ based on our null (usually that $\mu_1 - \mu_2$ = 0 but doesn't have to be)

Homogeneity of Variance

We assume homogeneity of variance for independent samples *t*-tests (i.e., even if samples come from different populations, these populations differ in mean but not variance), but in practice, this is unrealistic. We will most likely have two *different* variances

Standard error is easy to calculate when you only know one *SD* (e.g., one-sample *t*-tests). But we have two *SD*s, so unless they're the exact same, "we" need to calculate a weighted combination of the two.

How to calculate the Standard Error

Two main methods of calculating the *SE*:

1. **Student's *t*-test** is used when the variances of each sample are (roughly) the same. In many cases, slightly more powerful test but the **pooled variance** will be biased toward the sample with the larger sample size.
2. **Welch's *t*-test** if the variances are NOT (roughly) the same. Slightly less powerful.

Assumptions

Student's assumptions:

- Independence
- Normality of *Population* that each sample comes from
- Homogeneity of variance for *Population*

Welch's assumptions:

- Independence
- Normality of *Population* that each sample comes from

Variance Calculations

Student's

$$\hat{\sigma}_p^2 = \frac{(N_1 - 1)\hat{\sigma}_1^2 + (N_2 - 1)\hat{\sigma}_2^2}{N_1 + N_2 - 2}$$

$$\sqrt{\frac{(N_1 - 1)\hat{\sigma}_1^2 + (N_2 - 1)\hat{\sigma}_2^2}{N_1 + N_2 - 2}}$$

SE_{diff} Calculations

$$SE_{diff} = \hat{\sigma}_D$$

Welch's

$$\hat{\sigma}_D = \sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}$$

Essentially, the squared SE_{mean} of each sample is added then you take the square root.

Student's

$$\hat{\sigma}_D = \sqrt{\frac{(N_1 - 1)\hat{\sigma}_1^2 + (N_2 - 1)\hat{\sigma}_2^2}{N_1 + N_2 - 2}} \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$$

What did we just do?

Final result is *SE* of the difference of independent means. Remember that standard deviations estimate population variability (i.e., variability of population scores). Taking the next step and dividing the *SD* by the square root of some variation of N (depending on the test equation) helps us estimate variability of means rather than scores.

Welch's

$$\hat{\sigma}_D = \sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}$$

Student's

$$\hat{\sigma}_D = \sqrt{\frac{(N_1 - 1)\hat{\sigma}_1^2 + (N_2 - 1)\hat{\sigma}_2^2}{N_1 + N_2 - 2}} \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$$

Note that in Student's calculation, the variance for the larger sample will be weighted more heavily because each individual variance is being multiplied by N-1. Meanwhile, in Welch's, the standard deviations are being *divided* by sample size, so smaller samples will be weighted more heavily.

$$\hat{\sigma}_D = \sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}$$

```
# Welch's SE: sqrt(SD1^2/n1 + SD2^2/n2)
# random values
sqrt(4^2/1000 + 7^2/30)
```

```
## [1] 1.284264
```

```
sqrt(7^2/30)
```

```
## [1] 1.278019
```

One large sample can't make up for a small sample with Welch's.

df for Welch's

$$df = \frac{\left[\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}\right]^2}{\frac{[\frac{\hat{\sigma}_1^2}{N1}]^2}{N_1 - 1} + \frac{[\frac{\hat{\sigma}_2^2}{N2}]^2}{N_2 - 1}}$$

Welch's *df* will likely not be a whole number. And while Welch's allows for violations of homogeneity (i.e., unequal variance), you'll still be punished for it with reduced *df* unless you have equal variance and equal sample size. This doesn't mean that Welch can't have a lower *p*-value than student, though. The df will suffer, but the SE_{diff} could be lower to counteract that.

df for Student's

$$df = N_1 + N_2 - 2$$

Unlike before when we subtracted by 1, we are now calculating two means, so we need to subtract by two to get the *df*

We calculate our test statistic with:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_D}$$

and then can find the probability of the absolute value of this test statistic (for two-tailed) or more extreme given the null is true.

Examples!

Research Question: Do Kid or Adult trick R treaters get more candy

```
##
##  Descriptive statistics by group
## group: Adult
##    vars  n  mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 35 14.09 4.78     14   14.03 4.45   5  23    18 0.12    -0.93 0.81
## ------------------------------------------------------------
## group: Kid
##    vars   n mean   sd median trimmed  mad min max range  skew kurtosis   se
## X1    1 500 13.3 1.28     13   13.33 1.48   8  16     8 -0.27     0.11 0.06
```

Calculate Pooled Variance SE_{diff} (for Student's *t*-test)

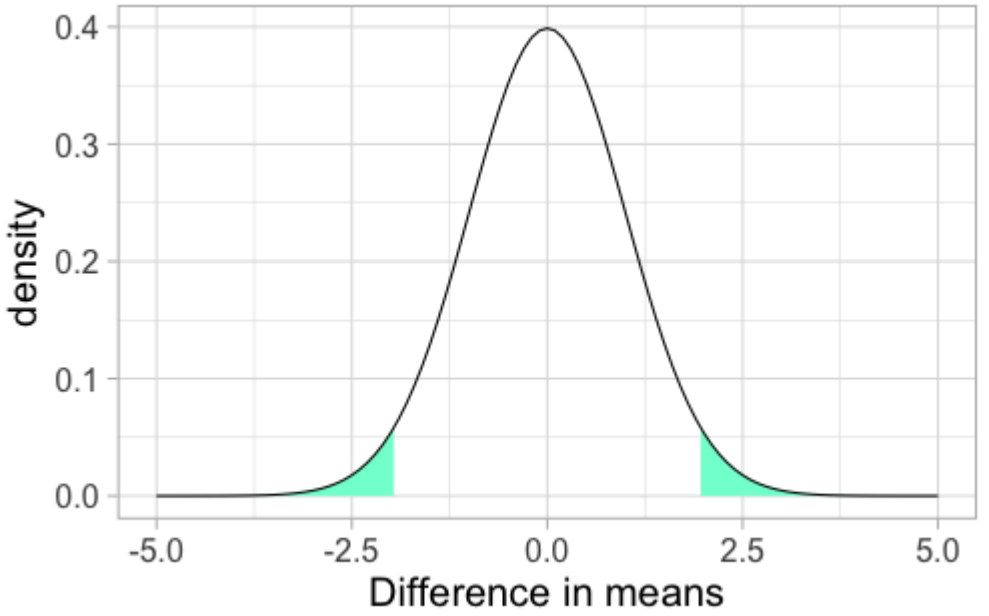
```
##
##  Descriptive statistics by group
## group: Adult
##    vars  n  mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 35 14.09 4.78     14   14.03 4.45   5  23    18 0.12    -0.93 0.81
## ------------------------------------------------------------
## group: Kid
##    vars   n mean   sd median trimmed  mad min max range  skew kurtosis   se
## X1    1 500 13.3 1.28     13   13.33 1.48   8  16     8 -0.27     0.11 0.06
```

$$\sigma_D = \sqrt{\frac{(500-1)1.28^2 + (35-1)4.78^2}{500+35-2}}\sqrt{\frac{1}{500}+\frac{1}{35}}$$

$$= 1.73\sqrt{\frac{1}{500}+\frac{1}{35}} = 0.30$$

$$df = 500 + 35 - 2$$

With this information, we can build a sampling distribution

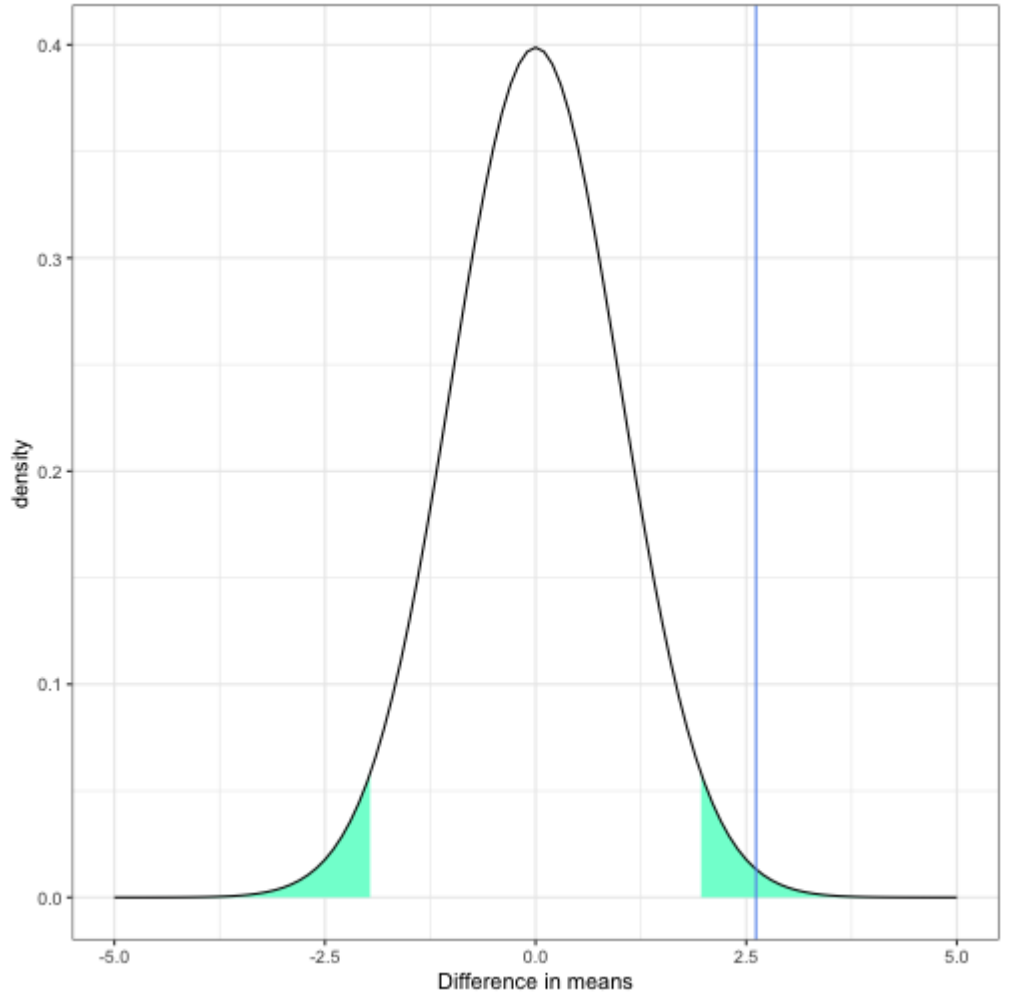


We have SE_{diff}. Now we can divide the mean differences by SE_{diff} to get student's *t* value.

$$t_{df} = \frac{\bar{X}_1 - \bar{X}_2}{SE_{meandifference}}$$

$$t_{df} = \frac{14.09 - 13.3}{0.302}$$

$$t(533) = 2.616$$



Our *t*-statistic exceeds the critical value cutoff, so our Student's *t*-test is significant.

Check your work!

Can use `pt` to confirm this.

```
2*pt(q=tvalue, df = 533, lower.tail = F)
```

```
## [1] 0.009148903
```

Can use `t.test` to extra confirm

```
t.test(Candy ~ Age, data = trickRtreat, var.equal = TRUE)
```

```
##
##     Two Sample t-test
##
## data:  Candy by Age
## t = 2.6126, df = 533, p-value = 0.009239
## alternative hypothesis: true difference in means between group Adult and group K
## 95 percent confidence interval:
##  0.1959301 1.3834985
## sample estimates:
## mean in group Adult   mean in group Kid
##            14.08571            13.29600
```

Confidence Intervals

Why stop there? We can also calculate confidence intervals of the mean difference

Confidence intervals are used to communicate the precision in how well our statistic estimates the parameter. As a reminder, they are grounded in frequentist probability: if we repeated our experiment or sampling infinitely, we would expect that 95% of our 95% confidence intervals would capture the true population parameter.

In an independent sample's *t*-test, we calculated three statistics, and so you can construct three different confidence intervals.Yay

Confidence interval around the difference in means

The most interpretable statistic is the difference in means -- this is the statistic you are testing using NHST.

$$CI_{\text{Difference}} = (\bar{X}_1 - \bar{X}_2) \pm \sigma_D(CV)$$

```
##
##  Descriptive statistics by group
## group: Adult
##    vars  n  mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 35 14.09 4.78     14   14.03 4.45   5  23    18 0.12    -0.93 0.81
## ------------------------------------------------------------
## group: Kid
##    vars   n mean   sd median trimmed  mad min max range  skew kurtosis   se
## X1    1 500 13.3 1.28     13   13.33 1.48   8  16     8 -0.27     0.11 0.06
```

Confidence interval around the difference in means

```
## [1] 3.022688
```

```
## [1] 1.964425
```

$$CI_{\text{Difference}} = (14.09 - 13.30) \pm .30(1.96)$$

$$[.20, 1.38]$$

Confidence interval doesn't include zero. Significant!

Confidence intervals around estimates of the mean

In addition to calculating precision of the estimate in difference in means, you may also want to calculate the precision of the mean estimates themselves.

In this case, you should use the standard deviation of the group sample as your estimate of population sd, instead of merging them.

$$\begin{aligned} CI_{\text{Mean}} &= \bar{X} \pm \sigma_M(CV) \\ &= \bar{X} \pm \frac{\hat{\sigma}}{\sqrt{N}}(CV) \end{aligned}$$

Adults

```
sd(trickRtreat$Candy[trickRtreat$A
```

```
## [1] 4.779684
```

```
qt(.975, df = 35-1)
```

```
## [1] 2.032245
```

```
14.09 + (4.78/sqrt(35)*2.03)
```

```
## [1] 15.73017
```

```
14.09 - (4.78/sqrt(35)*2.03)
```

```
## [1] 12.44983
```

$$14.09 \pm \frac{4.78}{\sqrt{35}}(2.03)$$

$$[12.45, 15.73]$$

Kids

```
sd(trickRtreat$Candy[trickRtreat$A
```

```
## [1] 1.278927
```

```
qt(.975, df = 500-1)
```

```
## [1] 1.964729
```

```
13.30 + (1.28/sqrt(500)*1.96)
```

```
## [1] 13.4122
```

```
13.30 - (1.28/sqrt(500)*1.96)
```

```
## [1] 13.1878
```

$$13.30 \pm \frac{1.28}{\sqrt{500}}(1.96)$$

$$[13.19, 13.41]$$

Do you believe it?

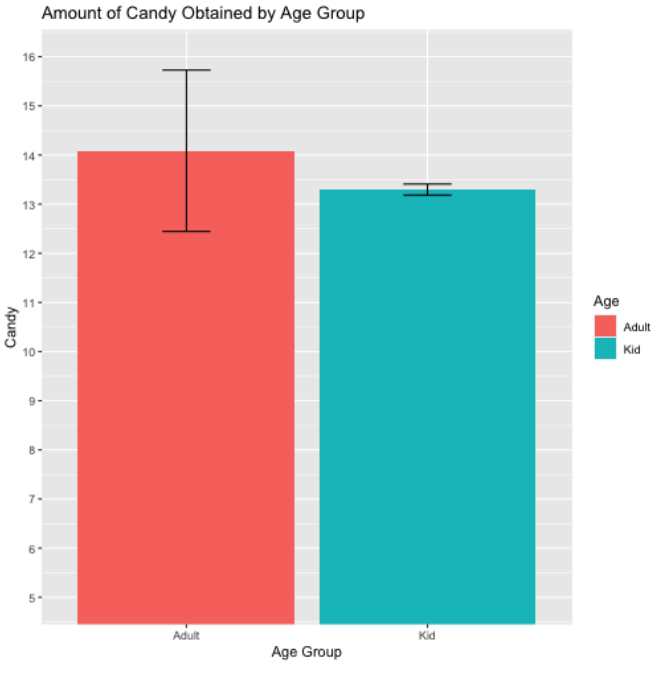
```
psych::describeBy(trickRtreat$Candy, group = trickRtreat$Age)
```

```
##
##  Descriptive statistics by group
## group: Adult
##    vars  n  mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 35 14.09 4.78     14   14.03 4.45   5  23    18 0.12    -0.93 0.81
## ------------------------------------------------------------
## group: Kid
##    vars   n mean   sd median trimmed  mad min max range  skew kurtosis   se
## X1    1 500 13.3 1.28     13   13.33 1.48   8  16     8 -0.27     0.11 0.06
```

SD seems to be pretty unequal, and the uneven sample size is going to exacerbate this issue. Student's *t* is pretty robust to homogeneity of variance violations when sample size is equal, but that's not the case here.

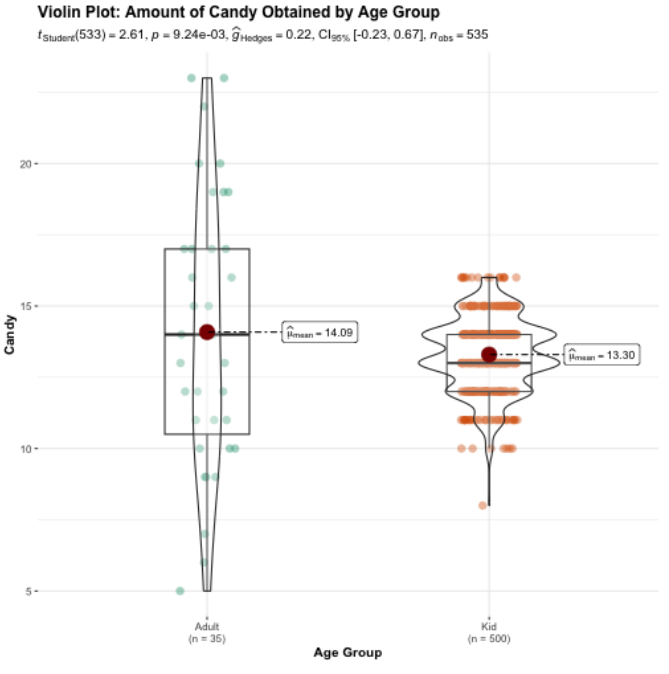
Plot it!

```
plot_data <- Rmisc::summarySE(trickRtreat, "Candy", "Age
plot_data %>%
  ggplot(aes(x = Age, y = Candy, fill=Age)) +
  geom_bar(stat = "identity") + ggtitle("Amount of Candy
  geom_errorbar(aes(ymin=Candy-ci, ymax=Candy+ci), width
  xlab("Age Group") +
  scale_y_continuous(breaks=seq(5,16,1)) +
  coord_cartesian( ylim = c(5,16))
```



Do better than bar graphs

```
library(ggstatsplot)
ggbetweenstats(
  data = trickRtreat,
  x = Age,
  y = Candy,
  xlab = "Age Group",
  ylab = "Candy",
  title = "Violin Plot: amount of
  results.subtitle = T,
  var.equal = T,
  bf.message = F,
  mean.size = 10,
  messages = F
)
```



Almost all the variability in Kids' candy scores are contained within Adults' interquartile (middle 50%) range. Yikes.

Welch's

This is where Welch's comes in handy (default)

```
t.test(Candy ~ Age, data = trickRtreat, var.equal = F) #var.equal = F us
```

```
##
##     Welch Two Sample t-test
##
## data:  Candy by Age
## t = 0.97503, df = 34.342, p-value = 0.3364
## alternative hypothesis: true difference in means between group Adult and group K
## 95 percent confidence interval:
##  -0.8556709  2.4350995
## sample estimates:
## mean in group Adult   mean in group Kid
##            14.08571            13.29600
```

With Welch's test, adults do not get significantly more candy than kids on Halloween. Sad. We can perform a significance test of homogeneity of variance as reassurance that the variances are indeed unequal.

Levene's Test for Homogeneity of Variance

Homogeneity of variance can be checked with Levene's procedure. It tests the null hypothesis that the variances for two or more groups are equal (or within sampling variability of each other):

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_A : \sigma_1^2 \neq \sigma_2^2$$

```
car::leveneTest(Candy~as.factor(Age), data = trickRtreat, center = "mean
```

```
## Levene's Test for Homogeneity of Variance (center = "mean")
##        Df F value    Pr(>F)
## group   1   301.3 < 2.2e-16 ***
##       533
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Levene's test gets more powerful as sample size increases. So unless your two variances are *exactly* equal to each other (and if they are, you don't need to do a test), your test will be "significant" with a large enough sample. Part of the analysis has to be an eyeball test -- is this "significant" because they are truly different, or because I have many subjects.

Don't need to base decision of Welch or Student based on a significance test. Utilize some of the tools we used earlier (looking at SDs, visualizations of spread and confidence intervals)

Normality

Finally, there's the assumption of normality. Specifically, this is the assumption that the population is normal -- if the population is normal, then our sampling distribution is **definitely** normal and we can use a *t*-distribution.

But even if the population is not normal, the Central Limit Theorem lets us assume our sampling distribution is normal because as N approaches infinity, the sampling distributions approaches normality. So we can be **pretty sure** the sampling distribution is normal.

One thing we can check -- the only distribution we actually have access to -- is the sample distribution. If this is normal, then we can again be pretty sure that our population distribution is normal, and thus our sampling distribution is normal too. But again, the sample distributions aren't required to be normally distributed.

Normality can be checked with a formal test: the Shapiro-Wilk test. The test statistic, *W*, has an expected value of 1 under the null hypothesis. Departures from normality reduce the size of *W*.

A statistically significant *W* is a signal that the sample distribution departs significantly from normal.

But...

- With large samples, even trivial departures will be statistically significant.
- With large samples, the sampling distribution of the mean(s) will approach normality, unless the data are very non-normally distributed.
- Visual inspection of the data can confirm if the latter is a problem.
- Visual inspection can also identify outliers that might influence the data.

```
shapiro.test(x = trickRtreat$Candy)
```

```
##
##     Shapiro-Wilk normality test
##
## data:  trickRtreat$Candy
## W = 0.88625, p-value < 2.2e-16
```

```
shapiro.test(x = trickRtreat$Candy[trickRtreat$Age == "Adult"])
```

```
##
##     Shapiro-Wilk normality test
##
## data:  trickRtreat$Candy[trickRtreat$Age == "Adult"]
## W = 0.9746, p-value = 0.5811
```

```
shapiro.test(x = trickRtreat$Candy[trickRtreat$Age == "Kid"])
```

```
##
##     Shapiro-Wilk normality test
##
## data:  trickRtreat$Candy[trickRtreat$Age == "Kid"]
## W = 0.9429, p-value = 6.062e-13
```

It's obvious that this isn't normal because the two groups have very different variabilities. It might seem surprising that Kids' significantly deviates from normalty, but we have a lot of Kids in the group.

What if I don't meet the normality assumption?

Non-parametric version of the independent samples *t*-test is the **Wilcoxon sum rank test**.

- Order all the data points by their outcome.
- For one of the groups, add up all the ranks. That's your test statistic, *W*.
- To build the sampling distribution, randomly shuffle the group labels and add up the ranks for your group of interest again. Repeat this process until you've calculated the rank sum for every possible group assignment.

```
wilcox.test(Candy~Age, data = trickRtreat)
```

```
##
##     Wilcoxon rank sum test with continuity correction
##
## data:  Candy by Age
## W = 9316, p-value = 0.5122
```

Effect Sizes: Do significant differences really make a difference?

Significance isn't a great proxy for meaningfulness or size of an effect. Many factors besides effect size goes into significance (provided the effect is not 0).

Cohen's d is a standardized mean difference and is one of the most common effect size estimate. Easy to understand and to compare across different experiments/studies.

$$\delta = \frac{\mu_1 - \mu_0}{\sigma} \approx d = \frac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma}_p}$$

Cohen's d is in the standard deviation (Z) metric.

Cohen's doesn't divide by *SE* so increasing sample size won't necessarily increase Cohen's D unless the effect grows or the variance decreases.

```
##
##  Descriptive statistics by group
## Age: Adult
##       vars  n  mean   sd median trimmed  mad min max range skew kurtosis   se
## Candy    1 35 14.09 4.78     14   14.03 4.45   5  23    18 0.12    -0.93 0.81
## ------------------------------------------------------------
## Age: Kid
##       vars   n mean   sd median trimmed  mad min max range  skew kurtosis   se
## Candy    1 500 13.3 1.28     13   13.33 1.48   8  16     8 -0.27     0.11 0.06
```

Kid

$$\bar{X}_1 = 13.3$$
$$\hat{\sigma}_1 = 1.28$$
$$N_1 = 500$$

Adult

$$\bar{X}_2 = 14.09$$
$$\hat{\sigma}_2 = 4.78$$
$$N_2 = 35$$

$$\hat{\sigma}_p = \sqrt{\frac{(35-1)4.78^2 + (500-1)1.28^2}{35+500-2}} = 1.73$$

$$d = \frac{14.09 - 13.3}{1.73} = 0.46$$

How do we interpret this? Is this a large effect?

Cohen (1988) suggests the following guidelines for interpreting the size of d:

- .2 = Small
- .5 = Medium
- .8 = Large

An aside, to calculate Cohen's D for a one-sample *t*-test:

$$d = \frac{\bar{X} - \mu}{\hat{\sigma}}$$

Cohen, J. (1988), Statistical power analysis for the behavioral sciences (2nd Ed.). Hillsdale: Lawrence Erlbaum.

Double check your work -- Student's *t*

```
library(lsr)
independentSamplesTTest(Candy ~ Age, data = trickRtreat, var.equal = T)
```

```
##
##    Student's independent samples t-test
##
## Outcome variable:   Candy
## Grouping variable:  Age
##
## Descriptive statistics:
##                Adult    Kid
##    mean     14.086 13.296
##    std dev.  4.780  1.279
##
## Hypotheses:
##    null:        population means equal for both groups
##    alternative: different population means in each group
##
## Test results:
##    t-statistic:  2.613
##    degrees of freedom:  533
##    p-value:  0.009
##
## Other information:
##    two-sided 95% confidence interval:  [0.196, 1.383]
##    estimated effect size (Cohen's d):  0.457
```

Double check your work -- Welch's *t*

```
library(lsr)
independentSamplesTTest(Candy ~ Age, data = trickRtreat, var.equal = F)
```

```
##
##    Welch's independent samples t-test
##
## Outcome variable:   Candy
## Grouping variable:  Age
##
## Descriptive statistics:
##                Adult    Kid
##    mean     14.086 13.296
##    std dev.  4.780  1.279
##
## Hypotheses:
##    null:        population means equal for both groups
##    alternative: different population means in each group
##
## Test results:
##    t-statistic:  0.975
##    degrees of freedom:  34.342
##    p-value:  0.336
##
## Other information:
##    two-sided 95% confidence interval:  [-0.856, 2.435]
##    estimated effect size (Cohen's d):  0.226
```