INFERENTIAL STATISTICS

Hypothesis Testing

Population vs Sample

Psycholinguistics is interested in how people process certain linguistic phenomena, but our experiments cannot collect data on every person and every instance, so instead we select a subset of people and instances to study

Population

- A complete set of individuals, events, utterances, etc.
- All German speakers
- All German sentences containing S (+s, –s) and K (+k, –k)

Estimate population by taking a sample

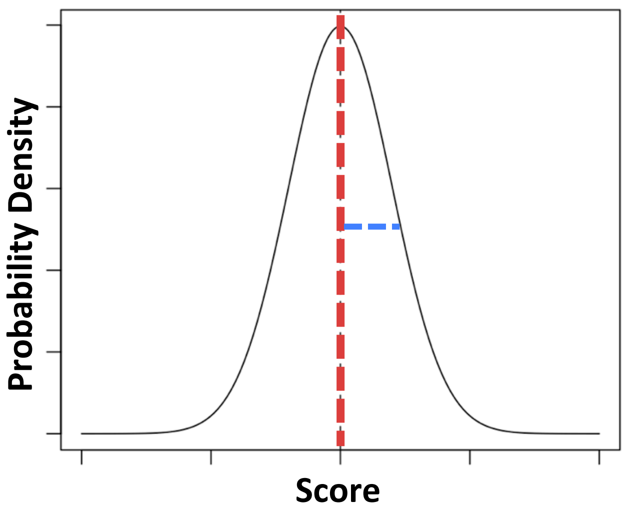
Sample

- A subset of a population
- Usually presumed to be randomly selected
- Central problem is uncertainty about how similar the sample is to the true population

Population vs Sample

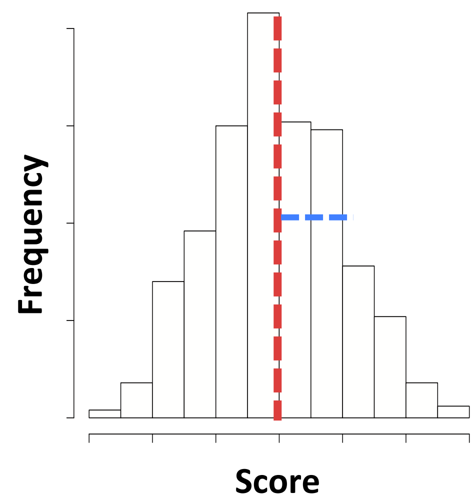
Psycholinguistics is interested in how people process certain linguistic phenomena, but our experiments cannot collect data on every person and every instance, so instead we select a subset of people and instances to study

Population



Estimate population by taking a sample

Sample



Inferential Statistics

Mean (***μ***)
Standard Deviation (***σ***)

Descriptive Statistics

Mean (***x̄***)
*Standard Deviation (**s**)*
etc.

Sampling in Psycholinguistics

In psycholinguistic experiments we sample from:

1) **Speakers**

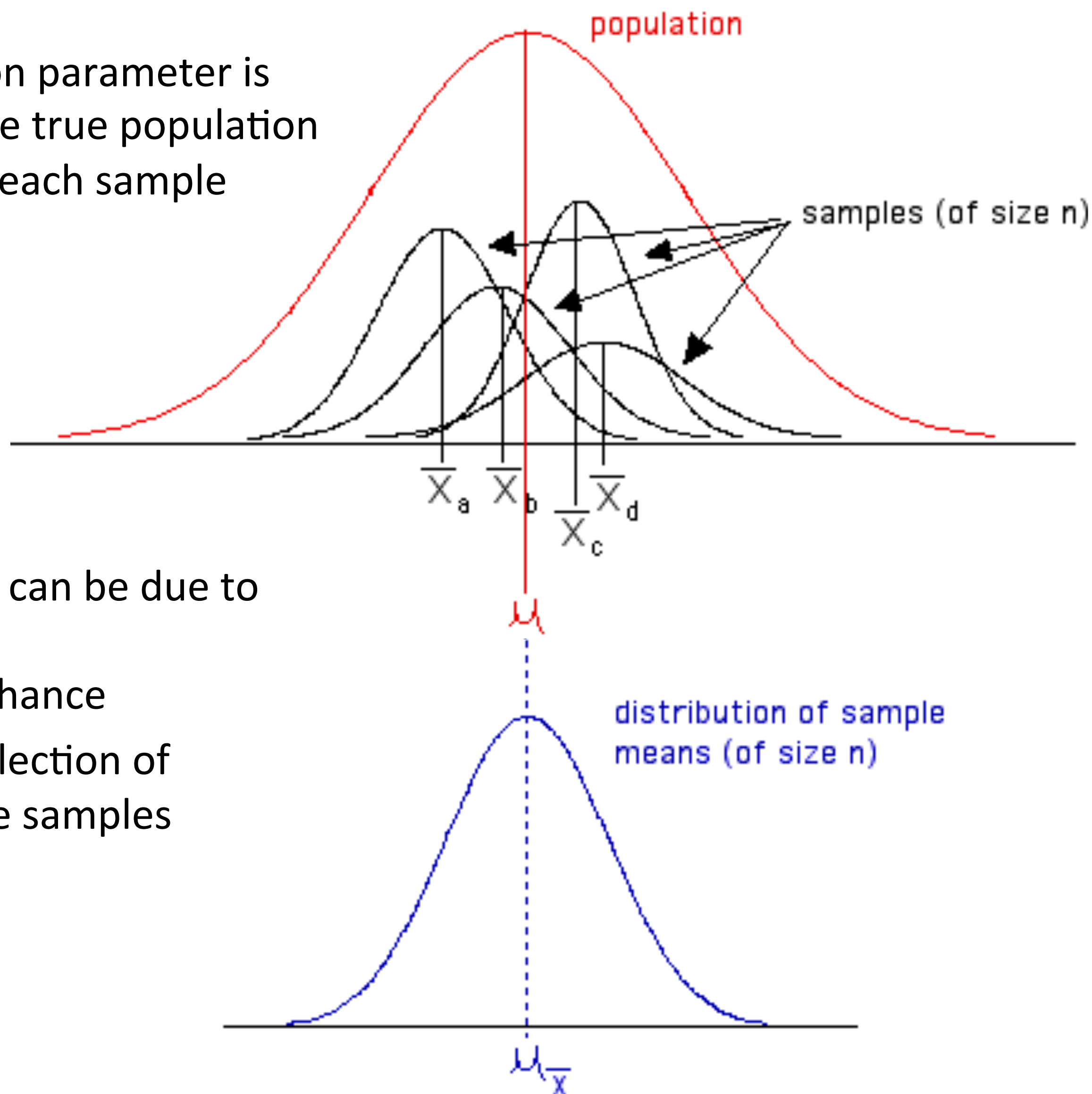
Analysis ***by Subjects***: We use inferential statistics to generalize the results to the population of all speakers of a language (not just the participants used in experiment)

2) **Language**

Analysis ***by Items***: We use inferential statistics to generalize the results to the entire collection of linguistic items displaying the phenomenon of interest (not just the items used in experiment)

Sampling Error and Bias

Each estimate of a population parameter is likely to be different from the true population parameter and different for each sample



→ These differences (error) can be due to

1. **Sampling Error** – Chance
2. **Sampling Bias** – Selection of non-representative samples

Hypothetical Data

Self-paced reading version of our study

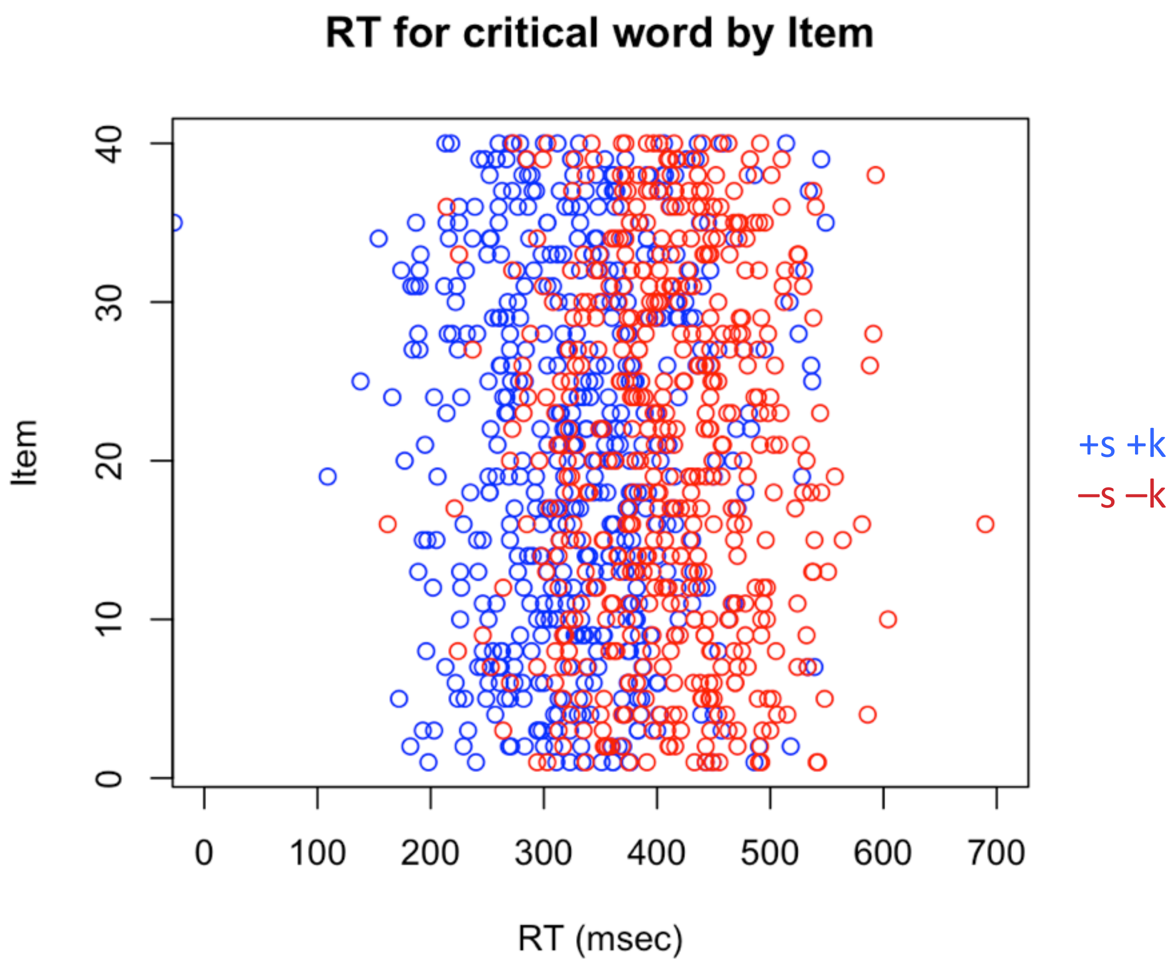
- 2 x 2 within-subjects design
  - Stereotype Consistency (+stereotypical, –stereotypical)
  - Speaker-specific Knowledge Consistency (+knowledge, –knowledge)

- DV: RT for critical word

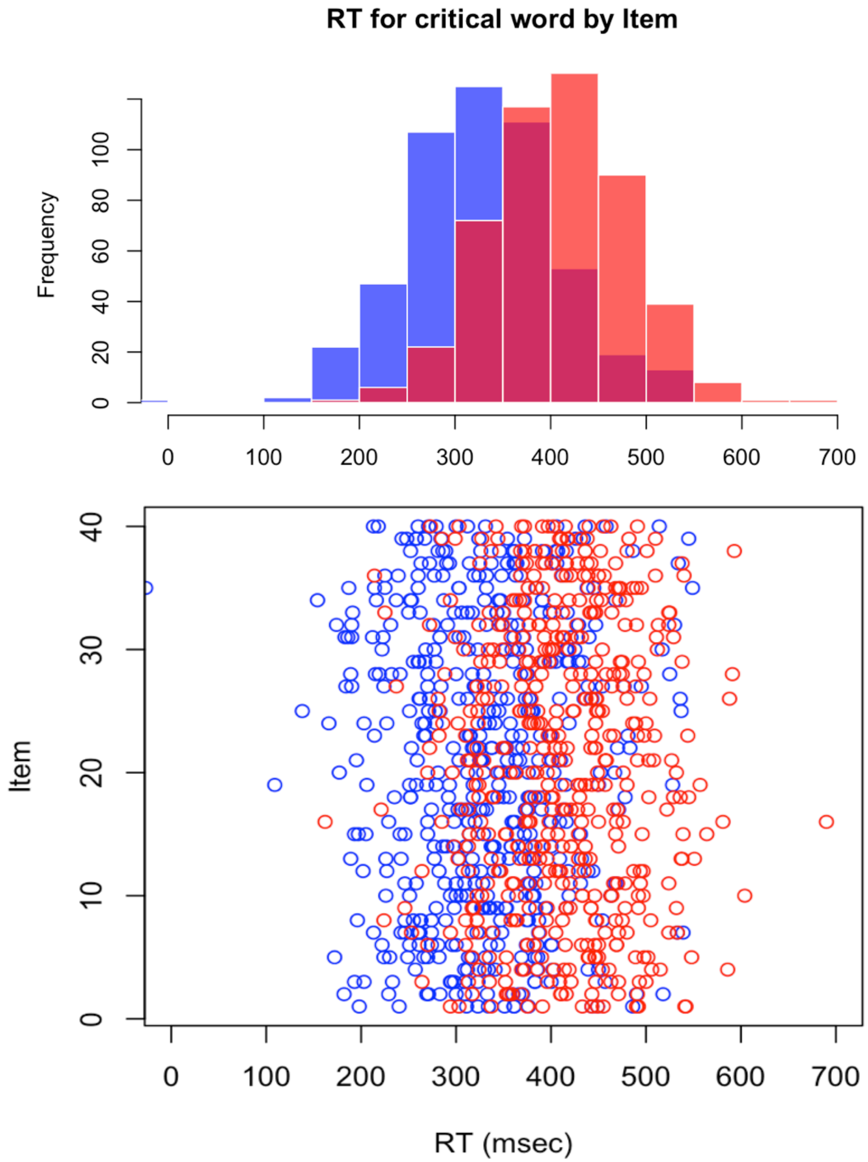
- Let’s say we will run 50 participants
- And each participant will see 40 items

→ How many data points will we have after running this experiment?
  - Per subject?
  - Per condition?

Hypothetical Data

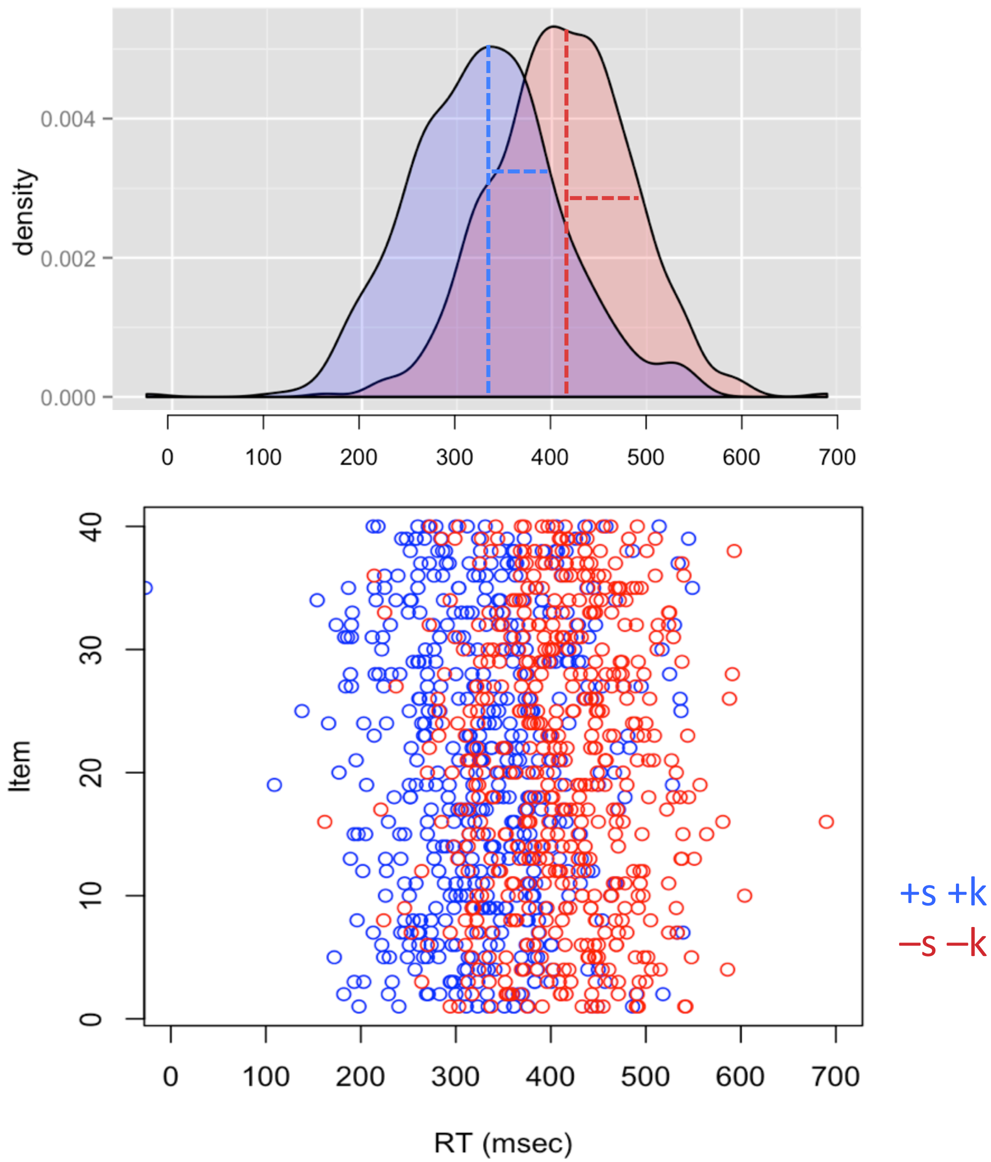


Hypothetical Data



+s +k
−s −k

Hypothetical Data



Are the differences between conditions significantly greater than what we would expect to see between any two samples drawn from the same population?

→ **Need statistical analysis**

How do we analyze our dataset?

1. Organize it
2. Summarize it
3. Apply statistical test

Hypothetical Data

Analysis by Subjects

Subj	+s +k	−s −k
1	312ms	325ms
2	365ms	356ms
3	200ms	224ms
4	324ms	388ms
5	356ms	412ms
6	326ms	378ms
7	279ms	299ms
…	…	…
50	323ms	340ms
$\bar{x}$	**320**	**350**
s	**48**	**55**

1. Organize: Aggregate data by subjects

2. Summarize: Descriptive statistics

Statistical Hypotheses Testing

To test whether a difference is significant, we first make the assumption that it is *not* (i.e., that the difference is just due to chance)

Null hypothesis (H_0)

$$\mu_{+s+k} = \mu_{-s-k} \Rightarrow \mu_{+s+k} - \mu_{-s-k} = 0$$

The research hypothesis, the outcome we predict, is that there is a *true* difference (i.e., that the difference is due to the manipulation)

Alternative hypothesis (H_1)

$a)\ \mu_{+s+k} \neq \mu_{-s-k} \Rightarrow \mu_{+s+k} - \mu_{-s-k} \neq 0$ Two-tailed hypothesis

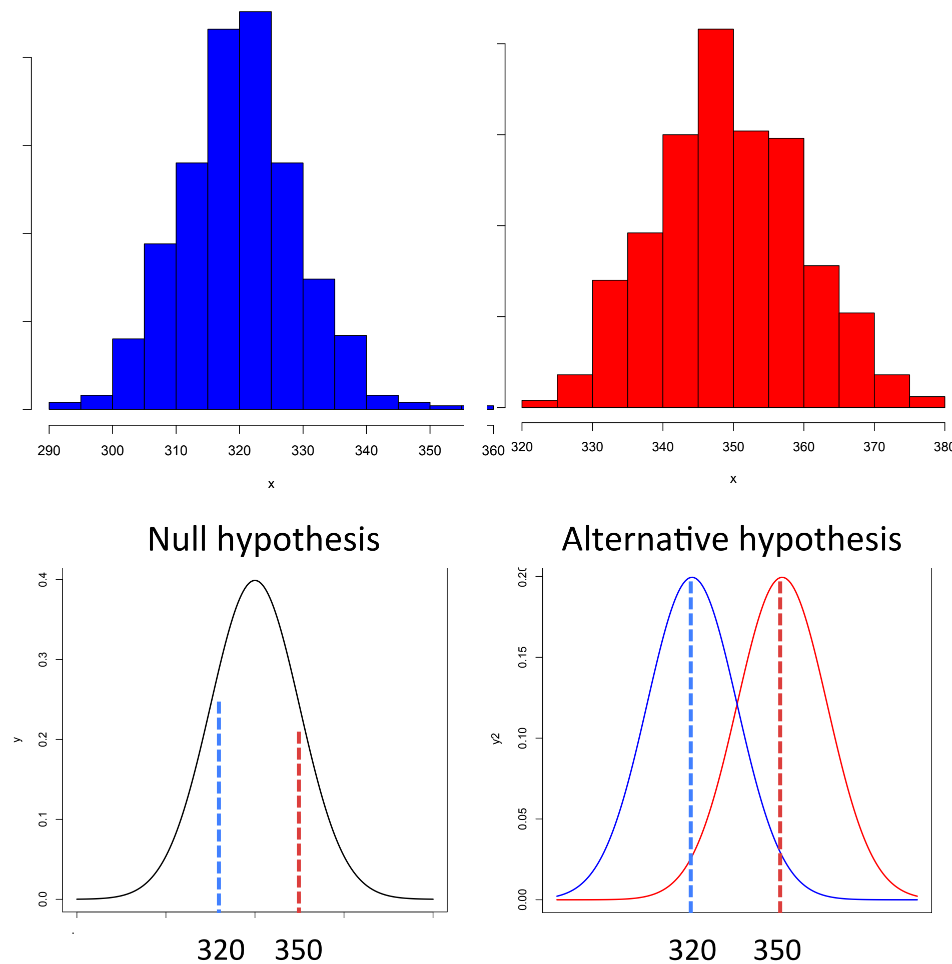
$b)\ \mu_{+s+k} - \mu_{-s-k} < 0$ One-tailed hypothesis

→ **Goal**: See if we can *reject* H_0, thereby supporting (BUT NOT PROVING) H_1

Hypothetical Data

3. Apply statistical test

Subj	+s +k	−s −k
1	312ms	325ms
2	365ms	356ms
3	200ms	224ms
4	324ms	388ms
5	356ms	412ms
6	326ms	378ms
7	279ms	299ms
…	…	…
50	323ms	340ms
$\bar{x}$	**320**	**350**
s	**48**	**55**



Hypothetical Data

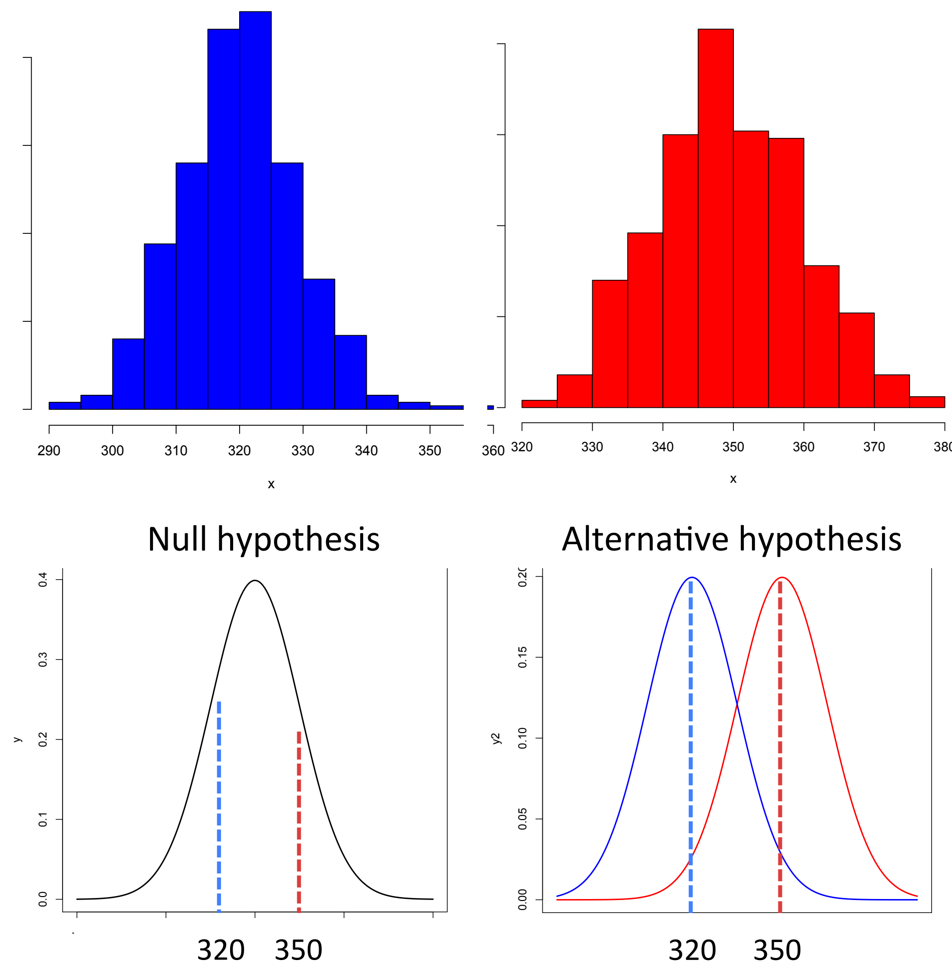
3. Apply statistical test

Statistical tests (e.g., *t*-test, ANOVA) tell us how likely it is to observe the difference, **assuming that the null hypothesis is true**

If this probability (p-value) is low (e.g., < 5%) then we can reject H_0

→ The difference is significant

- The same difference would likely be observed if we used different subjects (or items)
- Thus, the result generalizes to the population



Statistical Tests

What kind of statistical test should be used?

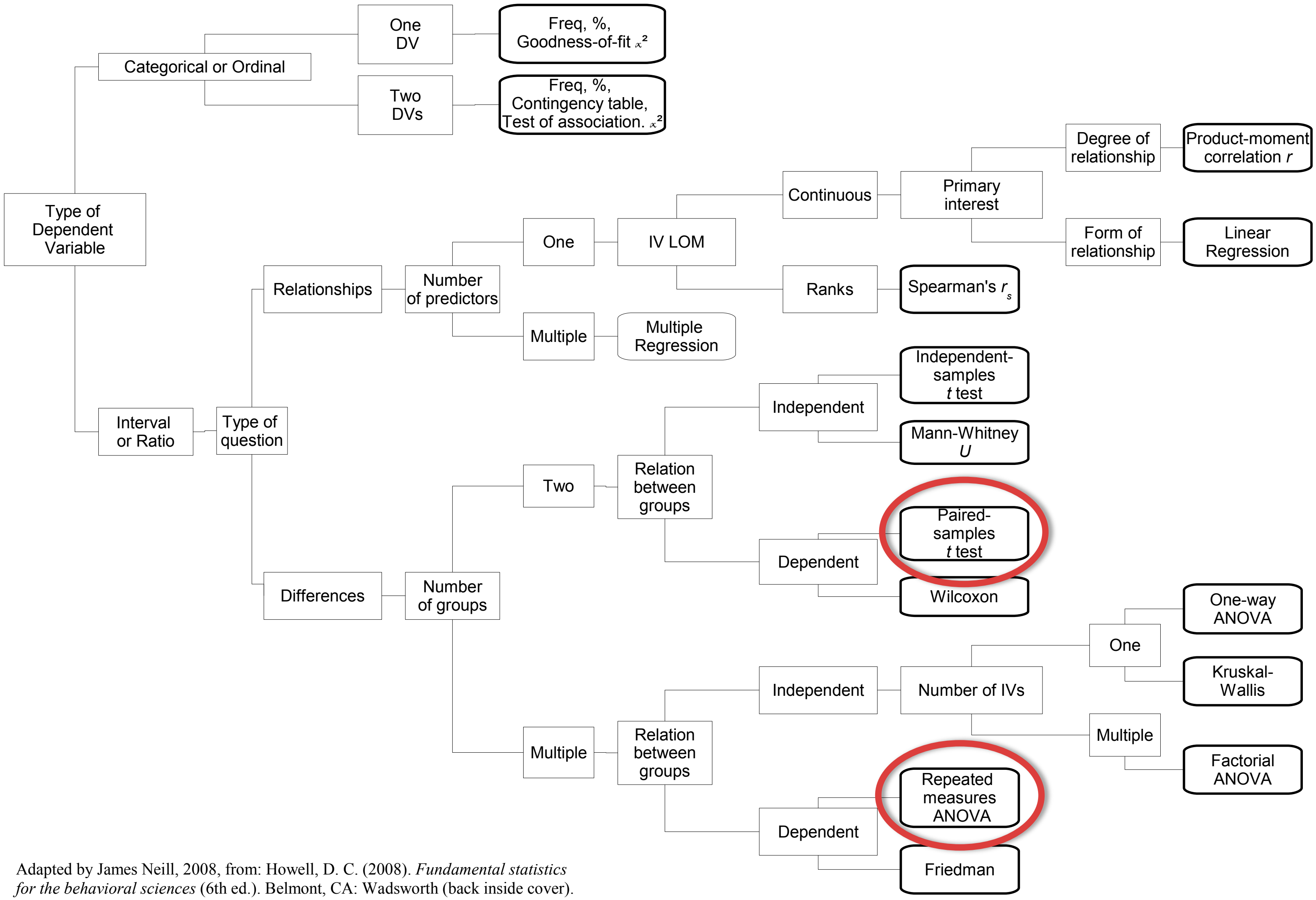
- *t*-test, ANOVA, χ^2-test, mixed effects models, etc.

Depends on

- Type of DV data (Continuous vs Categorical)
- Number of IVs in the design
- The assumed underlying distributions (normal, binomial, etc.)
- Whether design is between-subjects or within-subjects

In psycholinguistics, statistical analyses are performed both by-subjects (t_1, F_1) and by-items (t_2, F_2)

Decision Tree



Adapted by James Neill, 2008, from: Howell, D. C. (2008). *Fundamental statistics for the behavioral sciences* (6th ed.). Belmont, CA: Wadsworth (back inside cover).

Test types and what they tell us

Overview of two most common types of test in psychoinguistics

***t*-tests**

- Can only compare the effect of 1 IV with 2 levels
- Within-subjects: "Paired-samples" *t*-test (matched, related)
- Between-subjects: "Independent samples" *t*-test

Analysis of Variance (ANOVA)

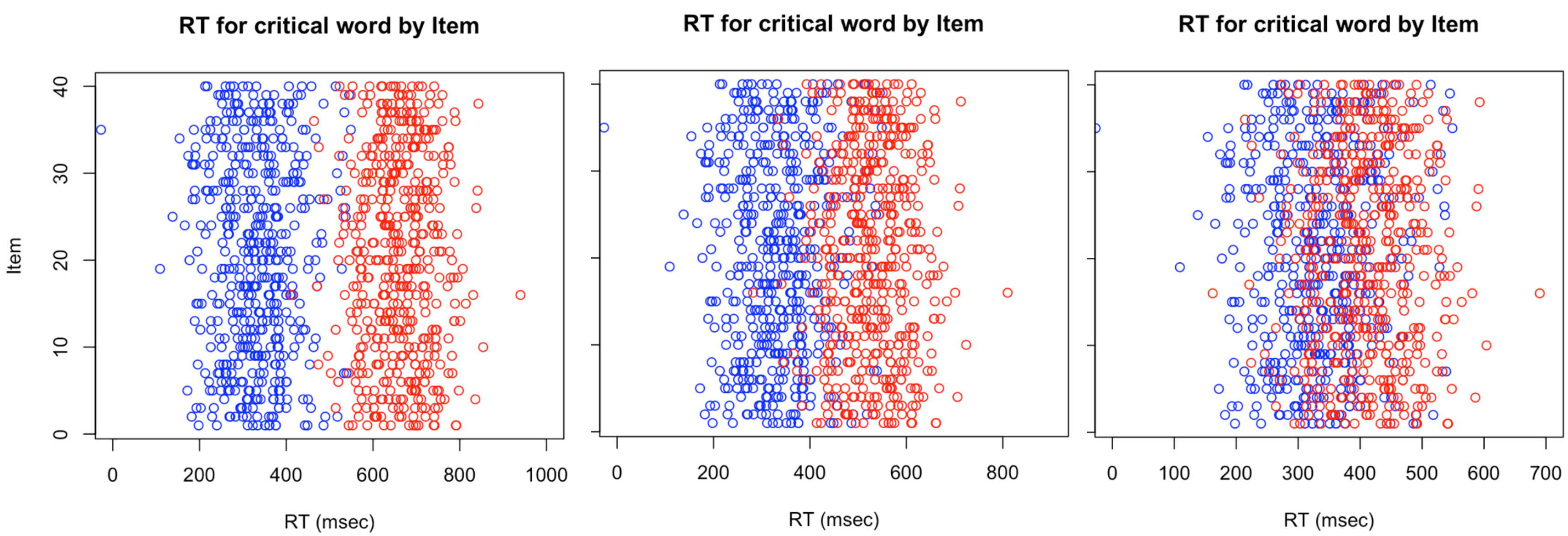
- Can do everything a *t*-test can do and more
- Can compare more than 2 IVs (factors) and more than 2 levels
- Can test for independent effects of each factor, even when factors have multiple levels
- Can test for interactions (i.e., relationships between factors)

Mixed effects models (LMER, GLMER)

- Next time

Variability Between and Within Conditions

Most statistical tests calculate a ratio that takes into account both sources of variability



Rule of thumb

If variability between conditions is large and variability within conditions (or groups) is relatively small, then the difference between conditions is likely to be significant

The *t*-test

Tests whether a difference between two means is significant

Simplified formula for paired-samples *t*-test (for repeated measures design)

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

Variability between conditions (signal)

Variability within conditions (noise)

Rule of thumb
If variability within conditions (or groups) is relatively small and variability between conditions is large, then the difference between conditions is likely to be significant

Hypothetical Data

3. Apply statistical test

Subj	+s +k	−s −k
1	312ms	325ms
2	365ms	356ms
3	200ms	224ms
4	324ms	388ms
5	356ms	412ms
6	326ms	378ms
7	279ms	299ms
...	...	...
50	323ms	340ms
$\bar{x}$	**320**	**350**
s	**48**	**55**

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

$$= \frac{-30}{\sqrt{(2304/50) + (3025/50)}} =$$

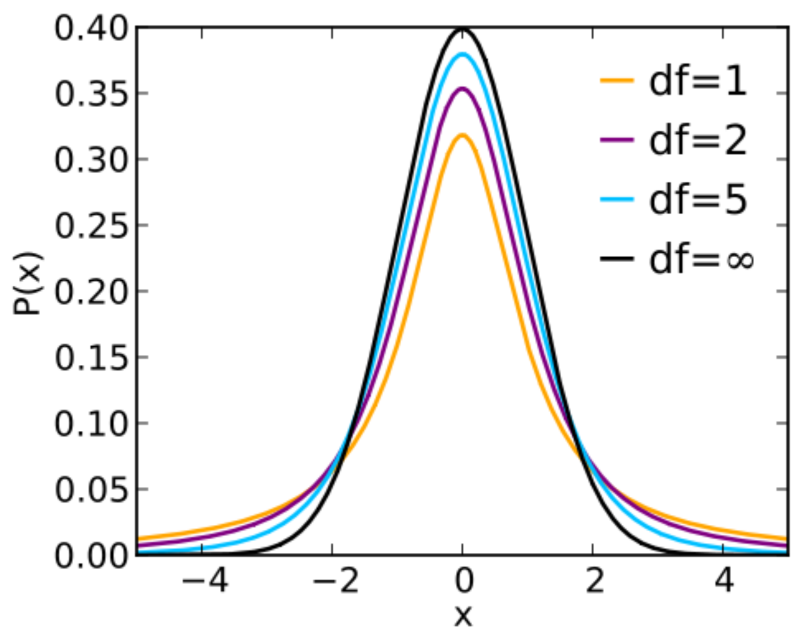
$$= \frac{-30}{\sqrt{46.08 + 106.58}} = -2.91$$

If the probability of observing ***t* = -2.91** is low, then we can reject the null hypothesis

How do we know if probability of *t* is low?

Look at the *t*-distribution (i.e., the distribution of all possible *t*-values if H_0 is true)

- *t*-values are normally distributed with $\mu = 0$
- Large or small values of *t* are rare, but values close to 0 are common



Exact shape of *t*-distribution depends on the number of **degrees of freedom (*df*)**

As ***df*** goes to infinity, the *t*-distribution converges to standard normal distribution ($\mu = 0,\ \sigma = 1$)

df = N - 1

df = 50 - 1 = 49

Degrees of Freedom (*df*)

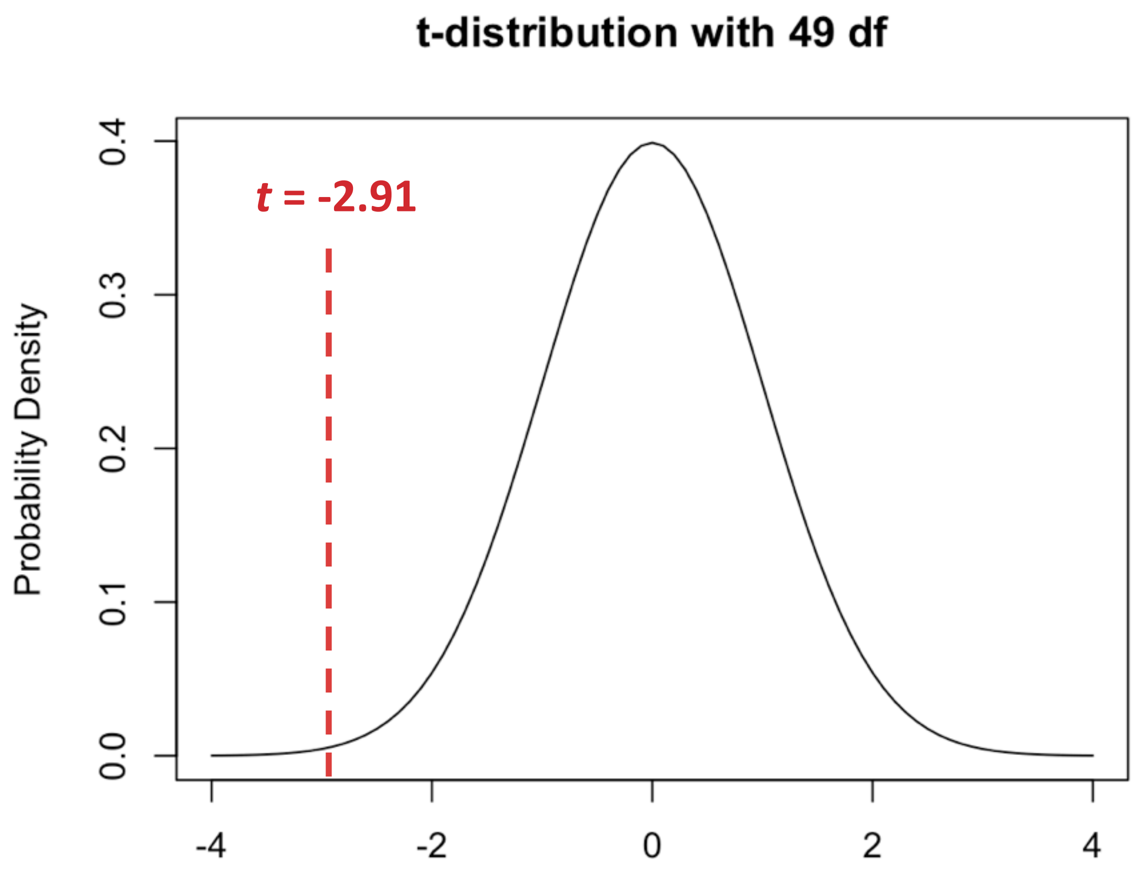
Defined as number of values in calculation of statistic that are "free to vary"

- Imagine you have four numbers (a, b, c, d) that must add up to X
- You are free to choose the first three numbers at random, but the fourth must be chosen so that it makes the total equal to X
- Thus, *df* = 3

→ ***df* provides a more accurate estimate of population parameters**

- Uncertainty about the true standard deviation requires us to qualify our beliefs (i.e., smaller sample sizes require us to be more conservative)
- As our sample size increases, *df* increases, which lowers the threshold for significance

t-distribution



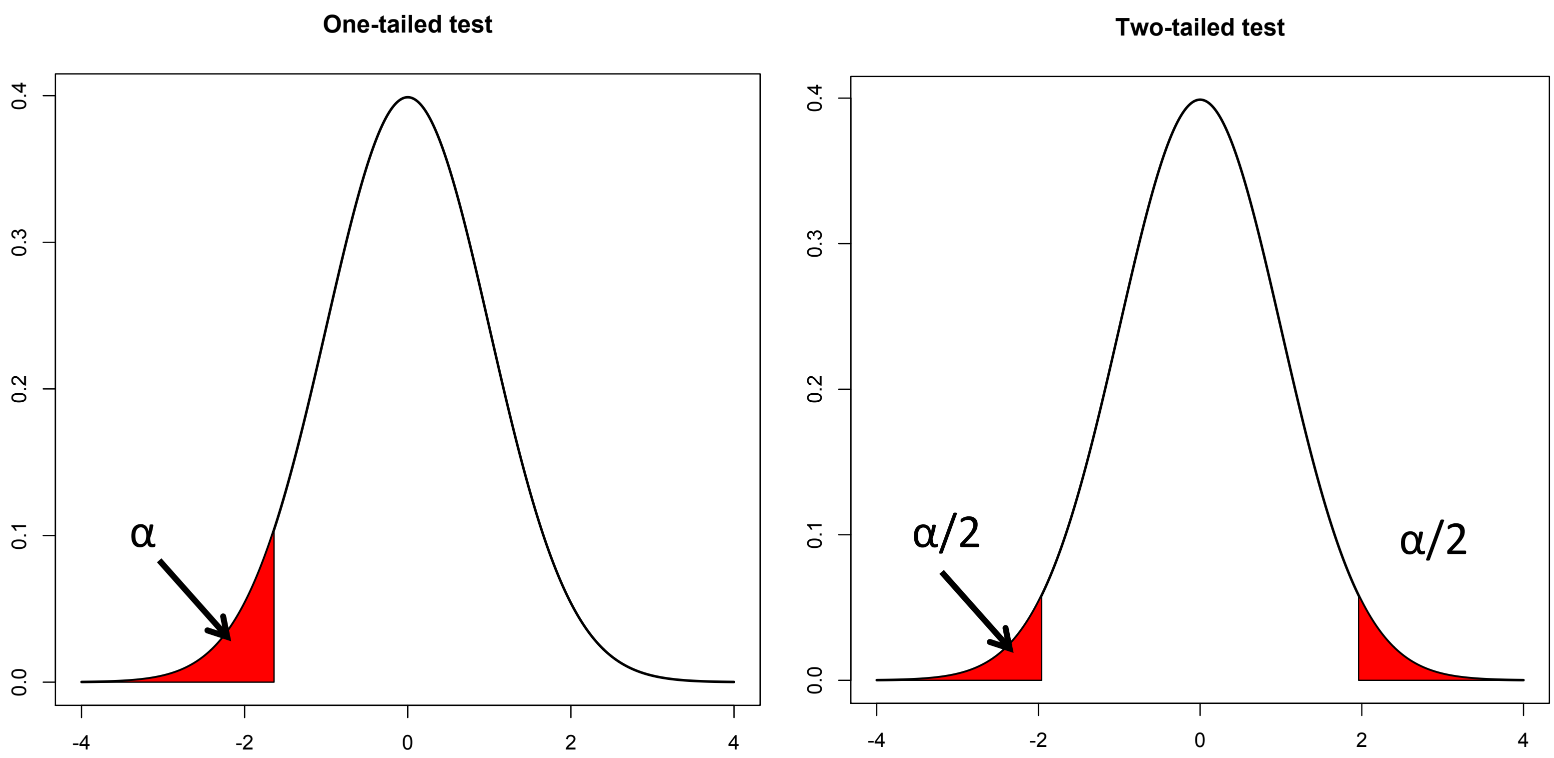
→ Can we reject H_0?

- Depends on the "critical value" of t (t_{crit})
- Which depends on our significance threshold (α)

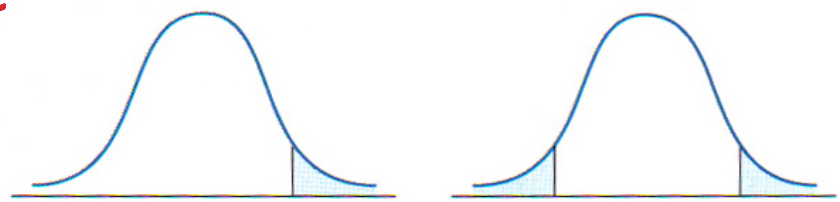
Significance Level: α

Alpha (α) – arbitrary cutoff value representing probability with which we are willing to reject H_0 when it is, in fact, true

α levels conventionally used: .05, .01



The *t*-table

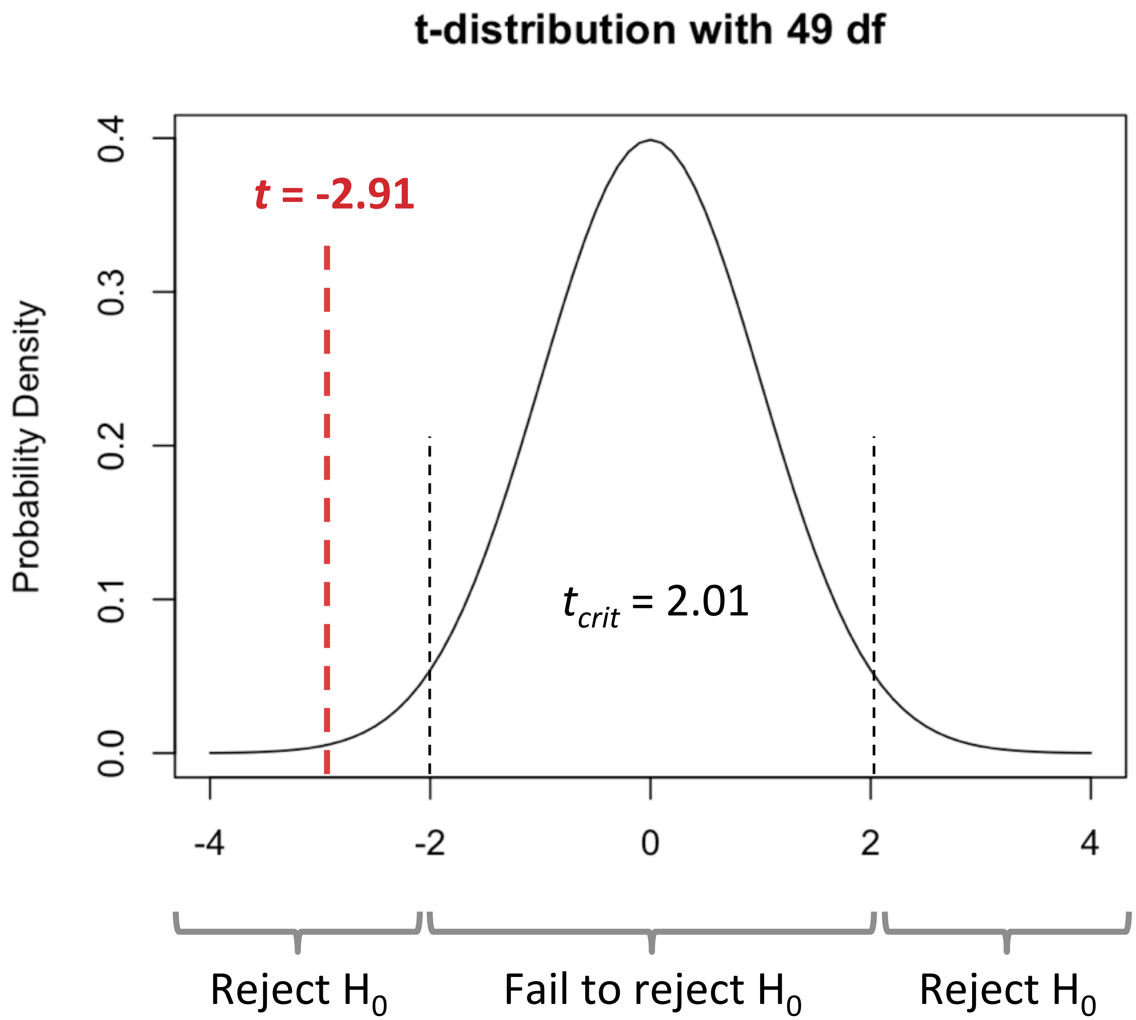


One tail
(either right or left)

Two tails
combined

	PROPORTION IN ONE TAIL					
	0.25	0.10	0.05	0.025	0.01	0.005
	PROPORTION IN TWO TAILS COMBINED					
df	0.50	0.20	0.10	0.05	0.02	0.01
1	1.000	3.078	6.314	12.706	31.821	63.657
2	0.816	1.886	2.920	4.303	6.965	9.925
3	0.765	1.638	2.353	3.182	4.541	5.841
4	0.741	1.533	2.132	2.776	3.747	4.604
5	0.727	1.476	2.015	2.571	3.365	4.032
6	0.718	1.440	1.943	2.447	3.143	3.707
7	0.711	1.415	1.895	2.365	2.998	3.499
8	0.706	1.397	1.860	2.306	2.896	3.355
9	0.703	1.383	1.833	2.262	2.821	3.250
10	0.700	1.372	1.812	2.228	2.764	3.169
11	0.697	1.363	1.796	2.201	2.718	3.106
12	0.695	1.356	1.782	2.179	2.681	3.055
13	0.694	1.350	1.771	2.160	2.650	3.012
14	0.692	1.345	1.761	2.145	2.624	2.977
15	0.691	1.341	1.753	2.131	2.602	2.947
16	0.690	1.337	1.746	2.120	2.583	2.921
17	0.689	1.333	1.740	2.110	2.567	2.898
18	0.688	1.330	1.734	2.101	2.552	2.878
40	0.681	1.303	1.684	2.021	2.423	2.704
60	0.679	1.296	1.671	2.000	2.390	2.660
120	0.677	1.289	1.658	1.980	2.358	2.617
∞	0.674	1.282	1.645	1.960	2.326	2.576

t-distribution



Paired-samples *t*-test

Calculating *t*-statistic and associated *p*-value in R (2-tail test)

```
# aggregated data (by subjects or by items) for contrast of interest
cw.data.2conds <- subset(HypData, Word.number==“cw” & (Cond==“ps.pk” | Cond==“ms.mk”))

t.test(Word.RT ~ Cond,                  # DV ~ IV (in this example, only 2 levels of Cond)
          data = cw.data.2conds,        # name of the dataset
          paired = TRUE,                # are the groups paired or not?
          mu = 0)                       # difference between means under H0

          Paired t-test

data:  Word.RT by Cond
t = -2.906, df = 49, p-value =  0.003382
alternative hypothesis: true difference in means is not equal to 0   # default = 2-tail
95 percent confidence interval:    # true difference between means is likely to fall
  -57.59908 -12.12092              # between -57.6 and -12.12
sample estimates:
mean of the differences
                 -34.86
```

Paired-samples *t*-test

Calculating *t*-statistic and associated *p*-value in R (1-tail test)

```
# aggregated data (by subjects or by items) for contrast of interest
cw.data.2conds <- subset(HypData, Word.number==“cw” & (Cond==“ps.pk” | Cond==“ms.mk”))

t.test(Word.RT ~ Cond,                  # DV ~ IV (in this example, only 2 levels of Cond)
          data = cw.data.2conds,        # name of the dataset
          paired = TRUE,                # are the groups paired or not?
          alternative = “less”,         # direction of H1
          mu = 0)                       # true difference between means under H0

          Paired t-test

data:  Word.RT by Cond
t = -2.906, df = 49, p-value =  0.001691
alternative hypothesis: true difference in means is less than 0   # 1-tail
95 percent confidence interval:    # true difference between means is likely to fall
      -Inf -15.88921               # between -Infinity and -15.9
sample estimates:
mean of the differences
                 -34.86
```

Effect Size

A statistical estimate of the magnitude of the effect that is relatively independent of sample size, and thus allows us to compare across studies

- Calculated after rejecting the null hypothesis (otherwise, it has no meaning)
- The larger the effect size, the fewer subjects needed to detect the effect
- Simple method: **r^2** (coefficient of determination, “r-squared”)

$$r^2 = \frac{t^2}{t^2 + df}$$

$$r^2 = \frac{-2.91^2}{-2.91^2 + 49}$$

$$r^2 = 0.1474$$

<u>By convention:</u>

0.01 small effect
0.09 medium effect
0.25 large effect

→ 14.8% of the variance in RT is predictable from the IV

Hypothetical Data

3. Apply statistical test

Subj	+s +k	−s −k
1	312ms	325ms
2	365ms	356ms
3	200ms	224ms
4	324ms	388ms
5	356ms	412ms
6	326ms	378ms
7	279ms	299ms
…	…	…
50	323ms	340ms
$\bar{x}$	**320**	**350**
s	**48**	**55**

1. Choose the alpha level (e.g., .05)
2. Calculate t score
3. Compare to t_{crit}

   If $|t| > t_\alpha$ then $p < .05$
   - The difference is significant
   - Reject H_0

   If $|t| \leq t_\alpha$ then $p \geq .05$
   - Null result
   - Fail to reject H_0

4. If H_0 rejected, calculate effect size

Reporting the Results

Report all of the following:

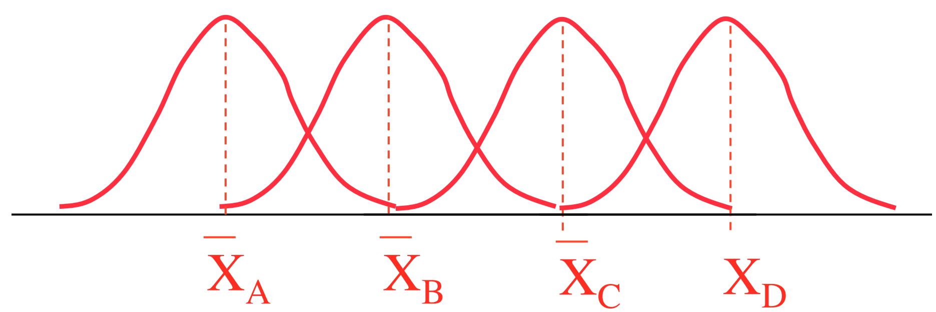
- The observed difference between conditions
- The specific kind of test (e.g., t-test)
- The computed statistic (e.g., t)
- Degrees of freedom for the test
- The p-value of the test
- The effect size (e.g., r^2)

"The mean response time for critical words in the +s+k condition was 30 ms faster than for the –s-k condition. A repeated measures *t*-test yielded a significant difference, $t(49) = -2.91$, $p < .01$, $r^2 = 0.147$."

Analysis of Variance (ANOVA)

Tests the effect of a categorical IV on a continuous DV

Test whether means of two or more conditions are significantly different from each other



$$F = \frac{\text{Variability between conditions (signal)}}{\text{Variability within conditions (noise)}}$$

Rule of thumb

If variability between conditions is large and variability within conditions (or groups) is relatively small, then the difference between conditions is likely to be significant

Analysis of Variance (ANOVA)

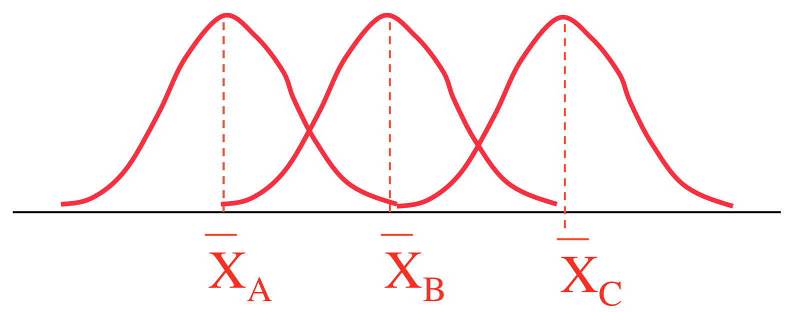
Categorized by number of IVs and whether groups are independent or dependent:

Type of ANOVA	Characteristics	With 2 levels, equivalent to:
One-way independent	• One IV with two or more levels • Each observation is independent	→ Independent samples *t*-test
One-way repeated measures	• One IV with two or more levels • Multiple observations from the same subjects	→ Paired-samples *t*-test
Two-way independent	• Two IVs (factors) with two or more levels • Each observation is independent	Factorial ANOVA
Two-way repeated measures	• Two factors with two or more levels • Multiple observations from the same subjects	
…		

- The term "way" describes number of IVs
- Factorial ANOVAs test how individual IVs and/or interactions affect the DV

Hypothesis Testing

e.g., One way ANOVA with 3 levels



Null hypothesis:

H_0: all the groups are equal

$\overline{X}_A = \overline{X}_B = \overline{X}_C$

Alternative hypothesis:

H_1: not all the groups are equal

$\overline{X}_A \neq \overline{X}_B \neq \overline{X}_C$ $\quad$ $\overline{X}_A \neq \overline{X}_B = \overline{X}_C$

$\overline{X}_A = \overline{X}_B \neq \overline{X}_C$ $\quad$ $\overline{X}_A = \overline{X}_C \neq \overline{X}_B$

→ **ANOVA tests H_0** Is there a difference between any one of these groups from another that is unlikely to be due to chance?

- If probability of *F* is low, we can reject H_0
- Then do follow-up pairwise comparisons

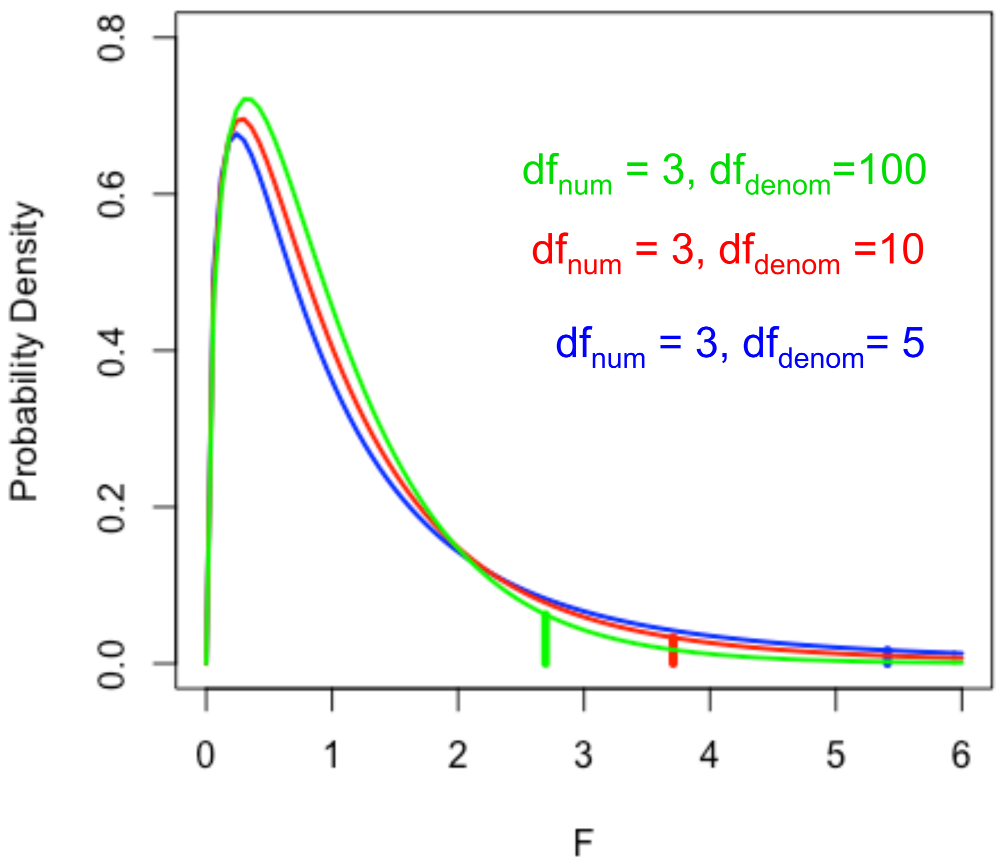
Test 1: A ≠ B $\quad$ Test 2: A ≠ C $\quad$ Test 3: B = C

p-values corrected for multiple comparisons (e.g., Bonferroni correction)

How do we know if probability of *F* is low?

Look at the *F*-distribution (i.e., the distribution of all possible *F*-values if H_0 is true)

- Shape of *F* distribution depends on two degrees of freedom (df_{num}, df_{denom})



$$F = \frac{\text{Variability between conditions (signal)}}{\text{Variability within conditions (noise)}}$$

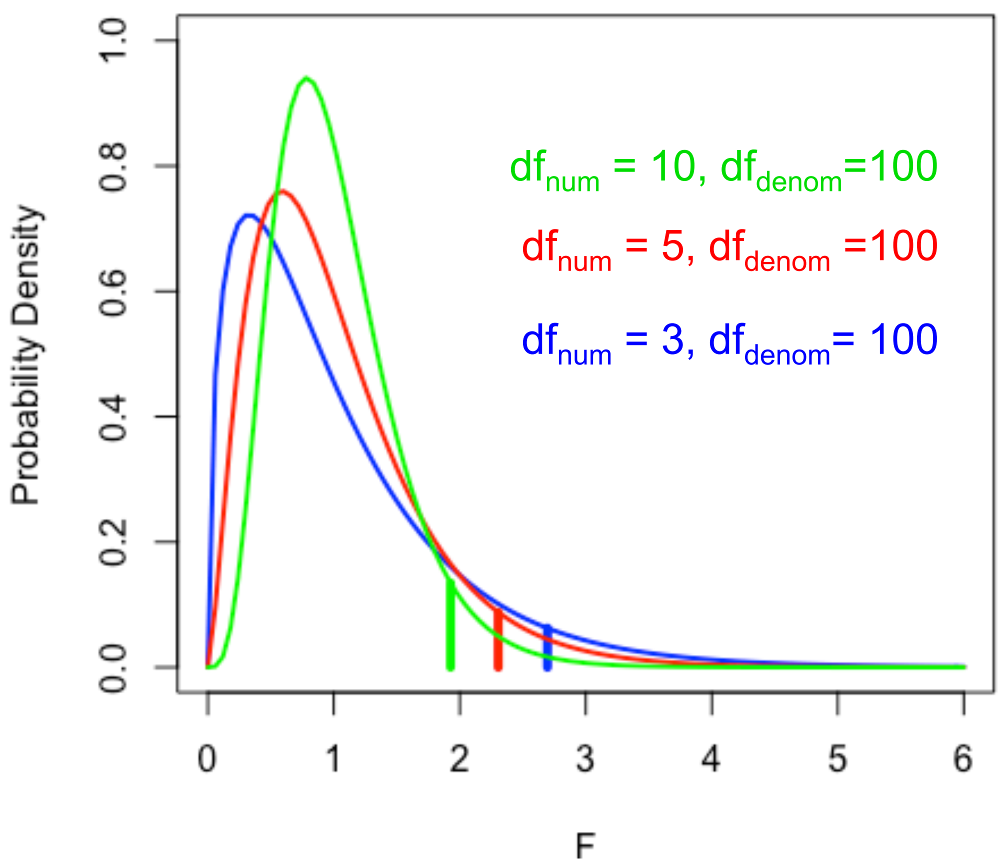
$$df_{num} = k - 1$$
$$k = \textit{number of conditions / groups}$$

$$df_{denom} = (n_{total} - 1)(k - 1)$$
$$n_{total} = \textit{total number of scores}$$

How do we know if probability of *F* is low?

Look at the *F*-distribution (i.e., the distribution of all possible *F*-values if H_0 is true)

- Shape of *F* distribution depends on two degrees of freedom (df_{num}, df_{denom})



$$F = \frac{\text{Variability between conditions (signal)}}{\text{Variability within conditions (noise)}}$$

$$df_{num} = k - 1$$
$$k = \textit{number of conditions / groups}$$

$$df_{denom} = (n_{total} - 1)(k - 1)$$
$$n_{total} = \textit{total number of scores}$$

Hypothetical Data

3. Apply statistical test

Subj	+s +k	–s –k
1	312ms	325ms
2	365ms	356ms
3	200ms	224ms
4	324ms	388ms
5	356ms	412ms
6	326ms	378ms
7	279ms	299ms
…	…	…
50	323ms	340ms
$\bar{x}$	**320**	**350**
s	**48**	**55**

1. Choose the alpha level (e.g., .05)
2. Calculate *F*-statistic, *p*-value, and effect size
3. If more than 2 levels within an IV, conduct follow-up pairwise comparisons with *p*-values corrected for multiple comparisons (e.g., Bonferroni correction)

Hypothetical Data

3. Apply statistical test

Subj	+s +k	−s −k
1	312ms	325ms
2	365ms	356ms
3	200ms	224ms
4	324ms	388ms
5	356ms	412ms
6	326ms	378ms
7	279ms	299ms
…	…	…
50	323ms	340ms
$\bar{x}$	**320**	**350**
s	**48**	**55**

$$F = \frac{\dfrac{\sum n_{in\ a\ group} * (\bar{x}_{group} - \bar{x}_G)^2}{k-1}}{\dfrac{\sum (x_i - \bar{x}_{group} - (\bar{x}_{subject} - \bar{x}_{group}))^2}{n_{total} - k - (n_{subjects} - 1)}}$$

$k = number\ of\ conditions\ /\ groups$

$n_{total} = total\ number\ of\ scores$

$\bar{x}_G = grand\ mean$

$n_{in\ a\ group} = number\ of\ scores\ in\ each\ group$

$\bar{x}_{group} = mean\ of\ the\ group\ the\ score\ comes\ from$

$x_i = individual\ score$

One-way repeated-measures ANOVA

Calculating *F*-statistic, *p*-value, and effect size (GES) in R using ezANOVA()

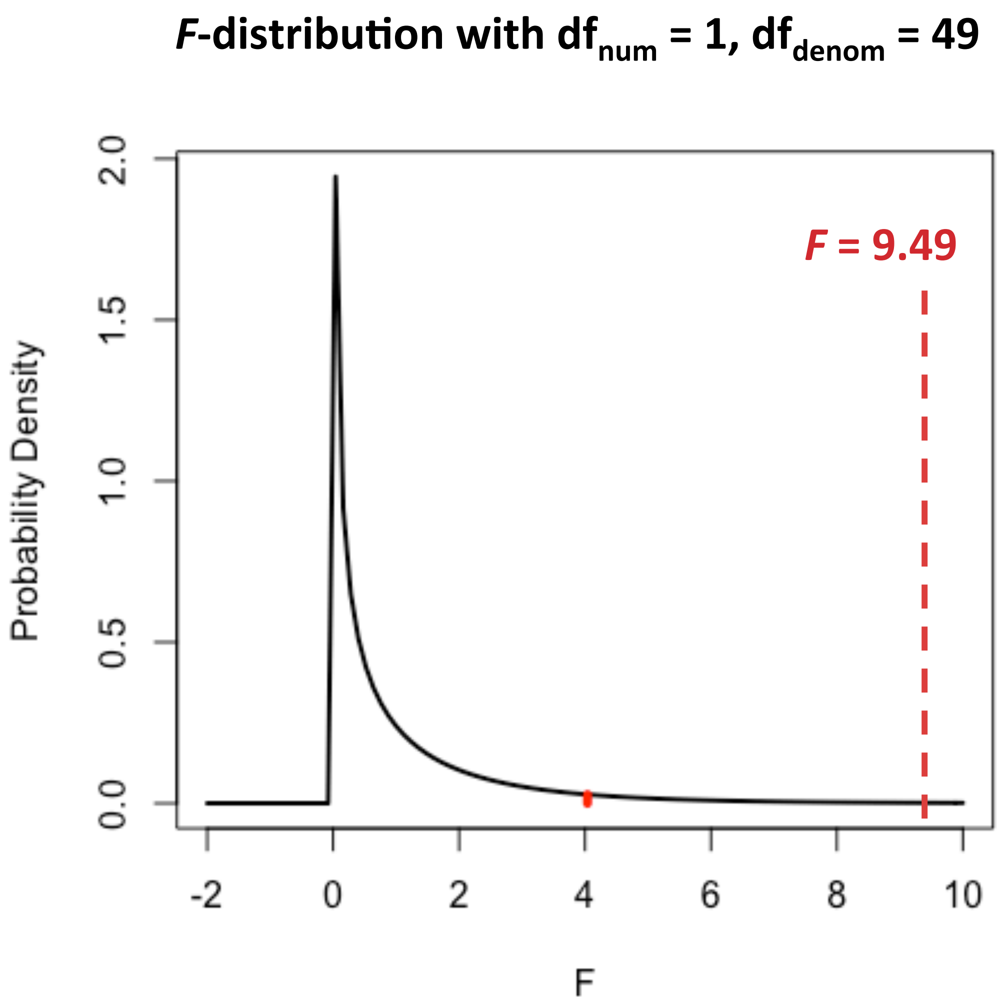
```
# aggregated data (by subjects or by items) for contrast of interest
cw.data.2conds <- subset(HypData, Word.number==“cw” & (Cond==“ps.pk” | Cond==“ms.mk”))

m1 <- ezANOVA(data = cw.data.2conds,  # name of dataframe
              dv = Word.RT,           # column name of DV
              wid = Subj,             # column name of within-group index
              within = Cond,          # cols for within-group IV(s)
              type=3)                 # type of sum of squares (SS)

m1                                    # output table of results

$ANOVA
  Effect DFn DFd        F           p p<.05       ges
2   Cond   1  49 9.491118 0.003382163     * 0.1055792
```

Generalized eta-squared (GES)

0.02	small effect
0.13	medium effect
0.26	large effect

→ The larger the effect size, the fewer subjects needed to detect the effect

One-way repeated-measures ANOVA



→ We can reject the null hypothesis

Paired-samples *t*-test

Calculating *t*-statistic and associated *p*-value in R

```
t.test(Word.RT ~ Cond,                # DV ~ IV
          data = cw.data.2conds,      # name of the dataset
          paired = TRUE,              # are the groups paired or not?
          mu = 0)                     # difference between means under H0
```

```
          Paired t-test

data:  Word.RT by Cond
t = -2.906, df = 49, p-value =  0.003382
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -57.59908 -12.12092
sample estimates:
mean of the differences
                 -34.86
```

$t^2 = F = 9.49$ ✓

One-way repeated-measures ANOVA

Calculate descriptive statistics using ezStats()

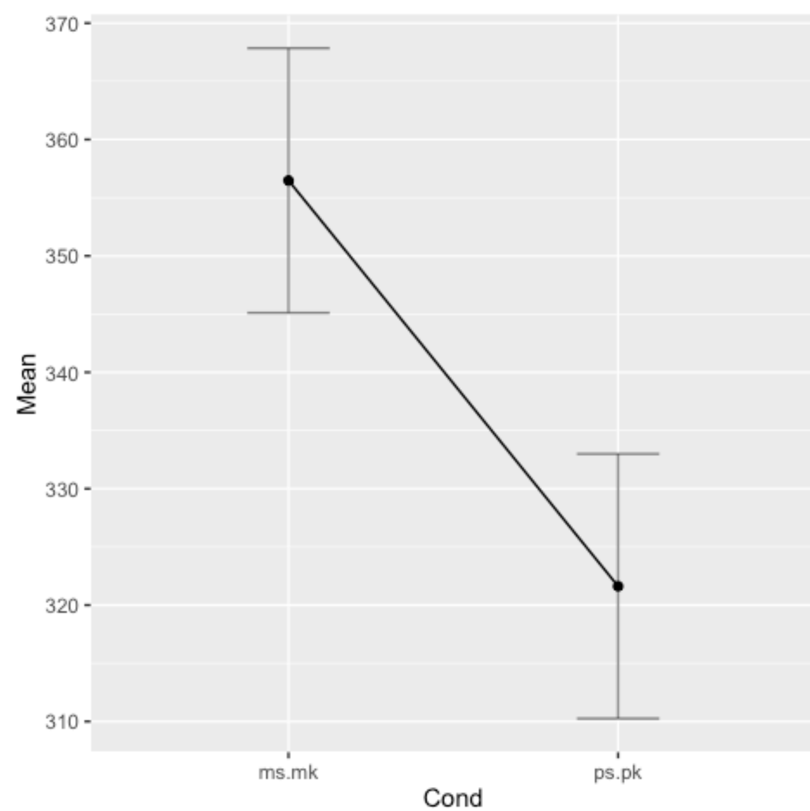
```
# get table of descriptive statistics
ezStats(data = cw.data.2conds,     # name of dataframe
              dv = Word.RT,        # column name of DV
              wid = Subj,          # column name of Subject index
              within = Cond,       # cols for within-subject IV(s)
              type=3)              # type of sum of squares (SS)
```

```
 Cond  N   Mean      SD     FLSD     # Fisher's Least Significant Difference (FLSD)
1 ms.mk 50 356.48 57.27493 22.73908
2 ps.pk 50 321.62 44.40762 22.73908
```

One-way repeated-measures ANOVA

Plot results using ezPlot()

```
# error bars represent FLSD
ezPlot(data = cw.data.2conds,
          dv = Word.RT,
          wid = Subj,
          within = Cond,
          type = 3,
          x = Cond) # IV on x-axis
```



Factorial ANOVA

Test whether individual factors and/or their interactions affect the DV

- Two or more factors (IVs)
  - Factors may be within or between
  - Overall design may be entirely within, entirely between, or mixed

- Multiple F-ratios are computed

  - One to test the **main effect** each factor:

    The effect of one IV on the DV, ignoring the effects of all other IVs

      Stereotypicality (+s, -s)

      Knowledge (+k, -k)

  - One to test each potential **interaction**:

    Whether the impact of one IV differs depending on value of another IV

      Stereotypicality x Knowledge (Stereotypicality:Knowledge)

Two-way repeated-measures ANOVA

3. Apply statistical test

Subj	+s +k	+s –k	–s +k	–s –k
1	312	333	341	325
2	365	389	368	356
3	200	221	277	224
4	324	312	365	388
5	356	367	399	412
6	326	399	387	378
7	279	295	296	299
…	…	…	…	…
50	323	333	339	340
x	**320**	**338**	**347**	**350**
s	**48**	**63**	**61**	**55**

1. Choose the alpha level (e.g., .05)
2. Calculate *F*-statistic*, p*-value*,* and effect size
3. If more than 2 levels within an IV, conduct follow-up pairwise comparisons with *p*-values corrected for multiple comparisons (e.g., Bonferroni correction)

Two-way repeated-measures ANOVA

Calculating *F*-statistic, *p*-value, and effect size in R using ezANOVA()

```
# aggregate data (by subjects or by items) for contrasts of interest
cw.data.allConds <- subset(HypData, Word.number==“cw”)

m2 <- ezANOVA(data = cw.data.allConds,       # name of dataframe
              dv = Word.RT,                  # column name of DV
              wid = Subj,                    # column name of within-group index
              within = .(Stereotypicality,   # cols for within-group IV(s)
                         Knowledge),
              type=3)                        # type of sum of squares (SS)

m2                                           # output table of results
$ANOVA
                      Effect DFn DFd          F            p p<.05          ges
2           Stereotypicality   1  49 19.6218320 5.293689e-05     * 0.0548009878
3                  Knowledge   1  49  1.1620173 2.863259e-01       0.0079323037
4 Stereotypicality:Knowledge   1  49  0.1098314 7.417495e-01       0.0003880361
```

Main effect of S

No effect of K

No S:K interaction

Two-way repeated-measures ANOVA

Calculate descriptive statistics using ezStats()

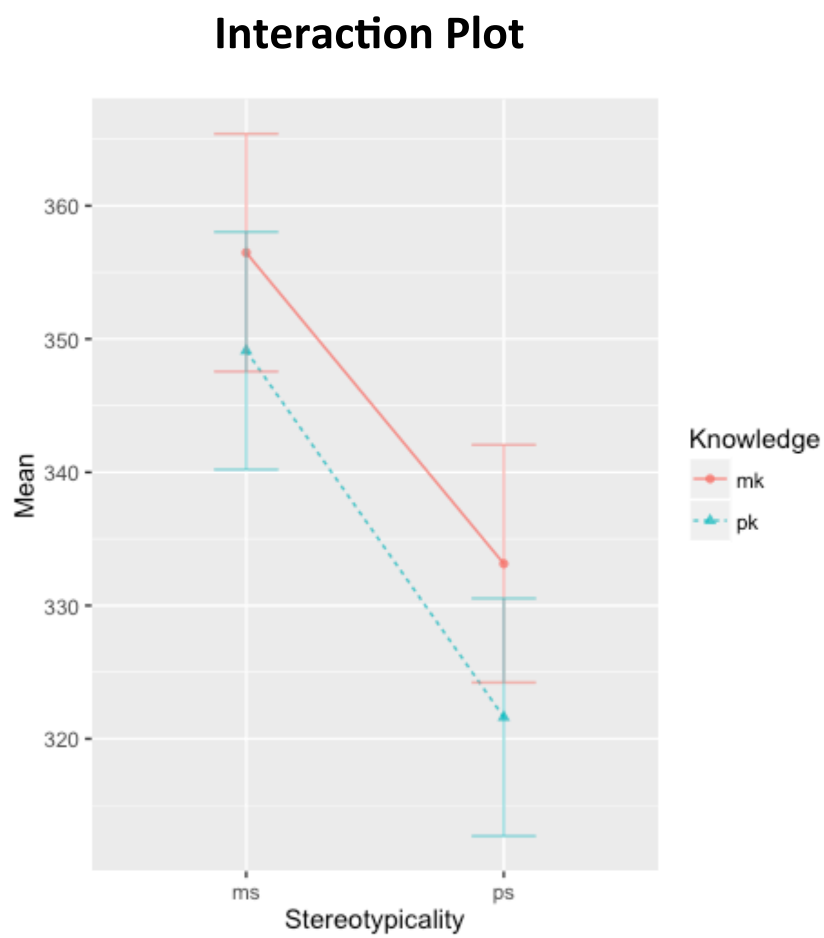
```
# get table of descriptive statistics
ezStats(data = data = cw.data.allConds,      # name of dataframe
               dv = Word.RT,                  # column name of DV
               wid = Subj,                    # column name of Subject index
               within = .(Stereotypicality,  # cols for within-subject IV(s)
                          Knowledge),
               type=3)                       # type of sum of squares (SS)
```

```
   Cond  N   Mean       SD     FLSD
1 ms.mk 50 356.48 57.27493 19.69711
2 ms.pk 50 349.12 56.48248 19.69711
3 ps.mk 50 333.14 54.12835 19.69711
4 ps.pk 50 321.62 44.40762 19.69711
```

Two-way repeated-measures ANOVA

Plot results using ezPlot()

```
# error bars represent FLSD
ezPlot(data = cw.data.allConds,
          dv = Word.RT,
          wid = Subj,
          within = .(Stereotypicality,
                     Knowledge),
          type=3,
          x = Stereotypicality,  # IV on x-axis
          split = Knowledge)     # IV by which
                                   to split data
                                   into colors
```



Reporting the Results

Report all of the following:

- The observed difference between conditions
- The specific kind of test (e.g., t-test)
- The computed statistic (e.g., t)
- Degrees of freedom for the test
- The p-value of the test
- The effect size (e.g., r^2)

"The mean response times for critical words were fastest in the +s+k condition, slowest for the -s-k condition, and intermediate for the +s-k and -s-k conditions (see Table 1). A 2x2 repeated measures ANOVA revealed a significant main effect of stereotypicality, $F_1(1,49) = 19.62$, $p < .001$, GES = 0.05, $F_2(1,39) = 8.33$, $p < .01$, GES = 0.07. No effect of speaker knowledge (*F*s < 2) or an interaction between stereotypicality or speaker knowledge (*F*s < 2) were found."

INTERPRETTING THE RESULTS

Possible outcomes of a 2x2 design

Interpreting the Results

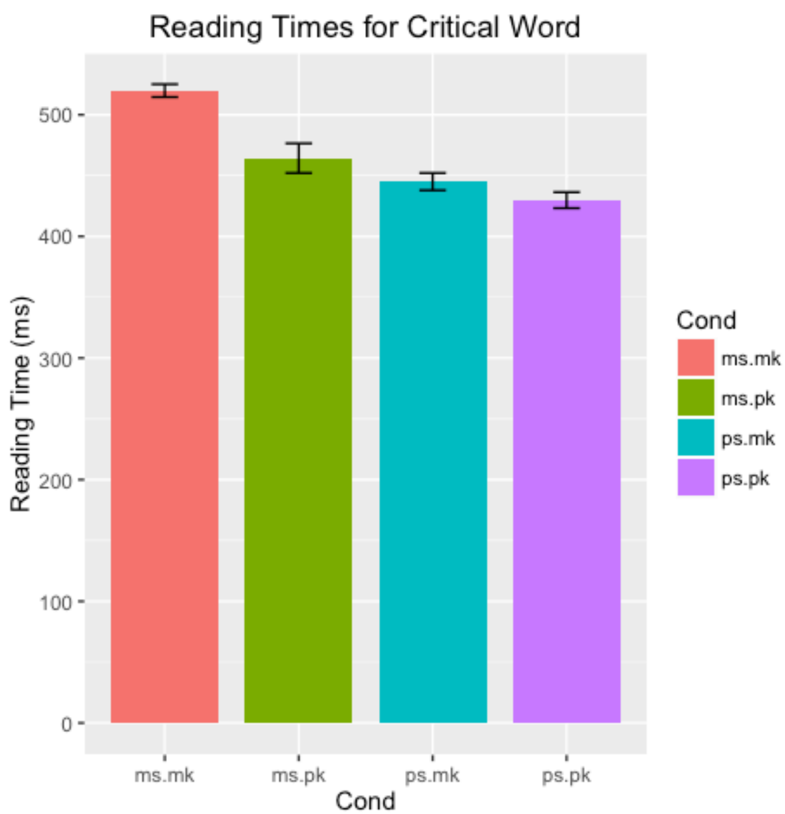
After collecting data, compare conditions to see whether IV(s) and/or their interaction had an effect on the DV

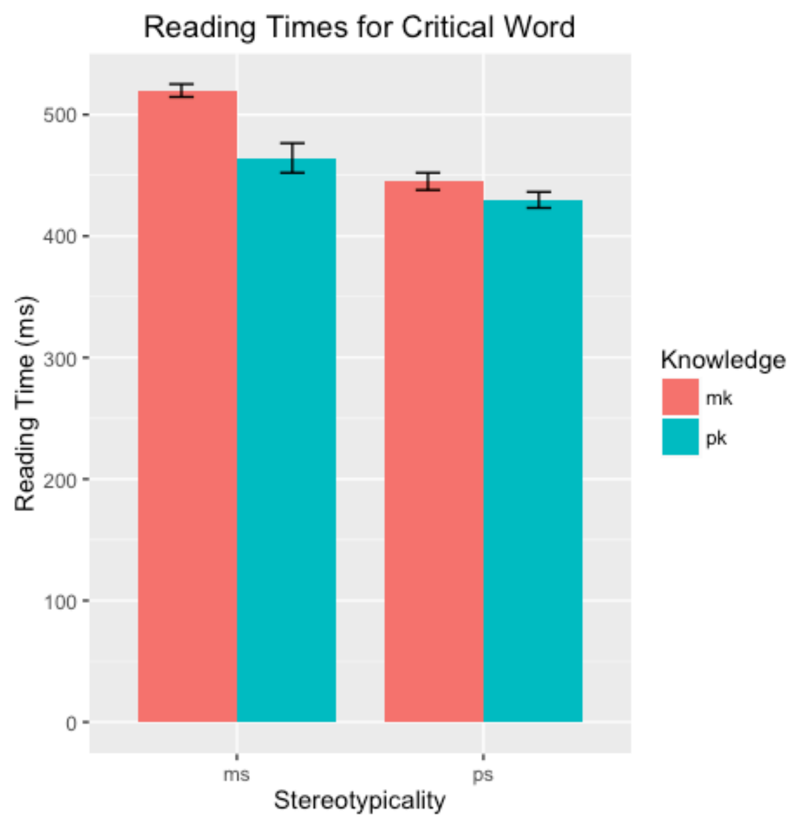
- **Graphs** – useful for identifying patterns and summarizing results
- **Statistical analyses** – tell us whether the results are likely to be "real"

Plotting the results (hypothetical data*)

Bar graph

- y-axis: DV, x-axis: IV levels
- Error bars usually represent Standard Error of the Mean (SEM))

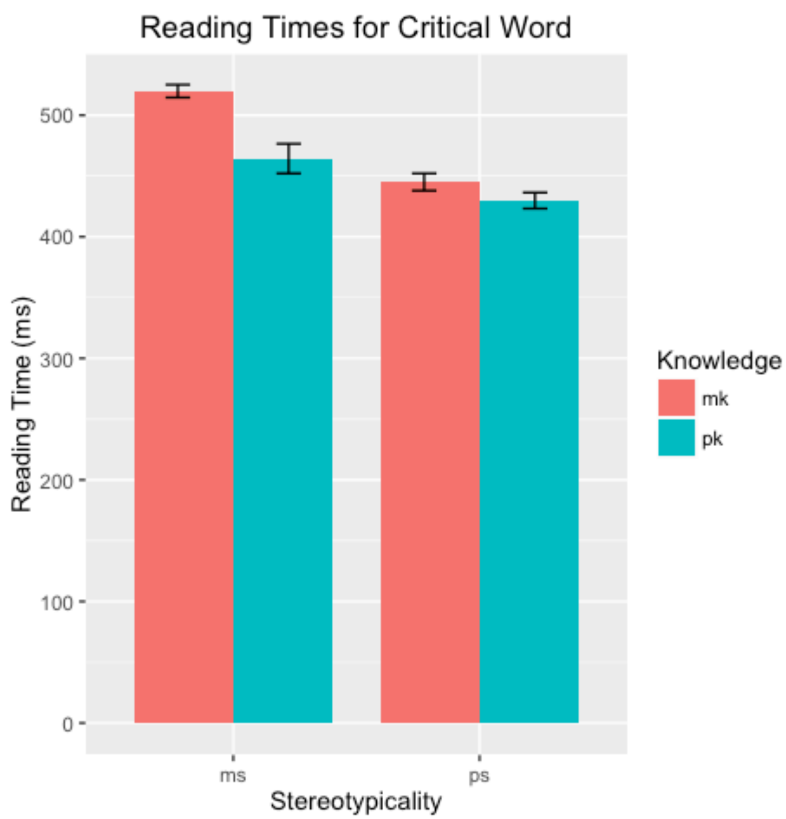




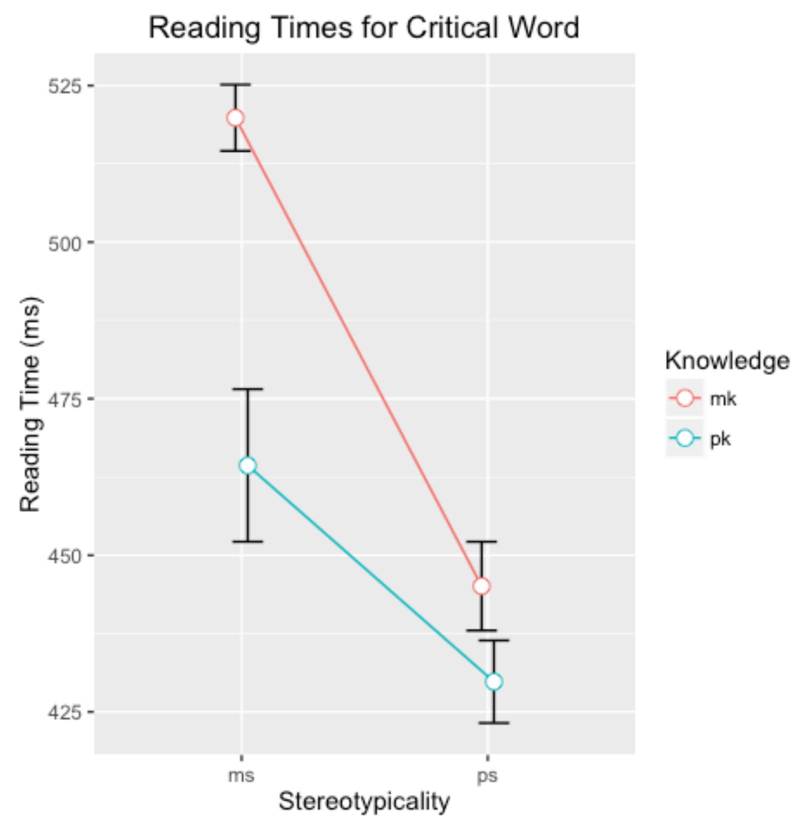
* Each of the following slides uses a different set of hypothetical data

Plotting the results (hypothetical data)

Bar graph

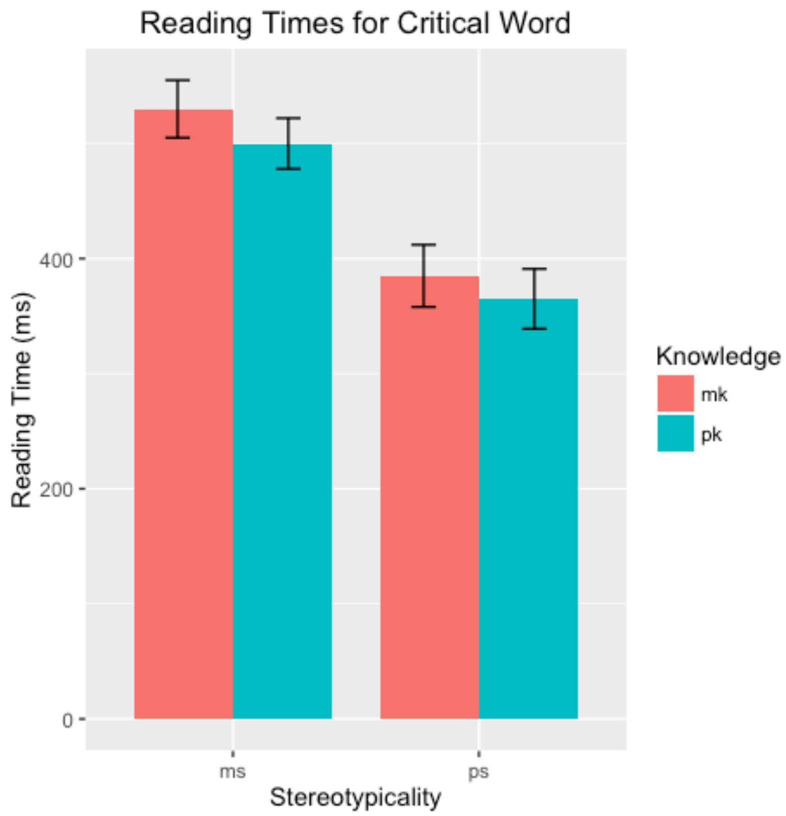


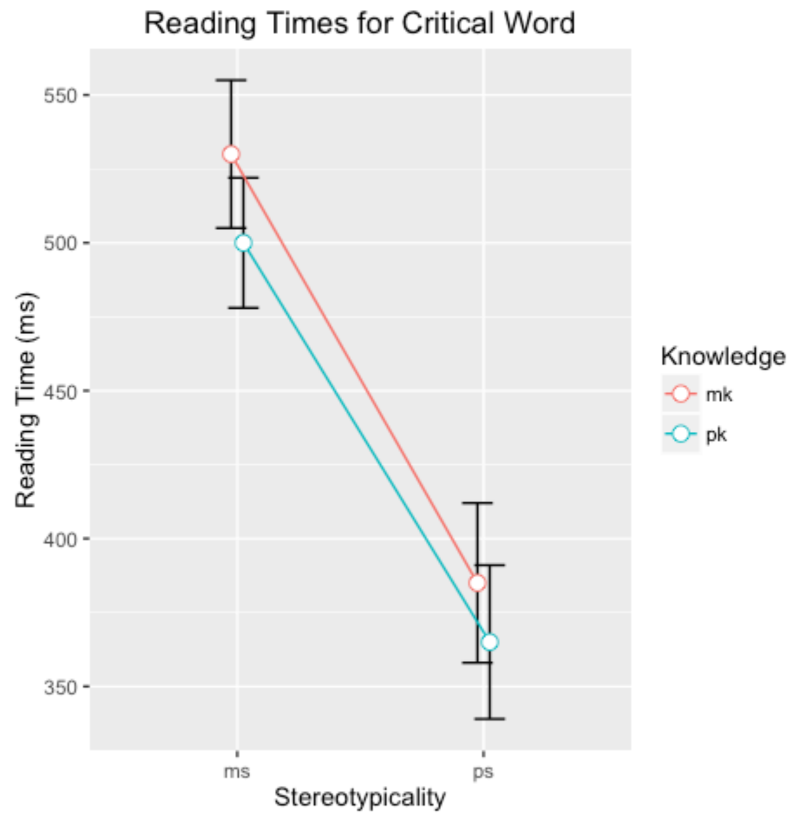
Line graph (Interaction plot)



Possible Outcomes of a 2x2 Design

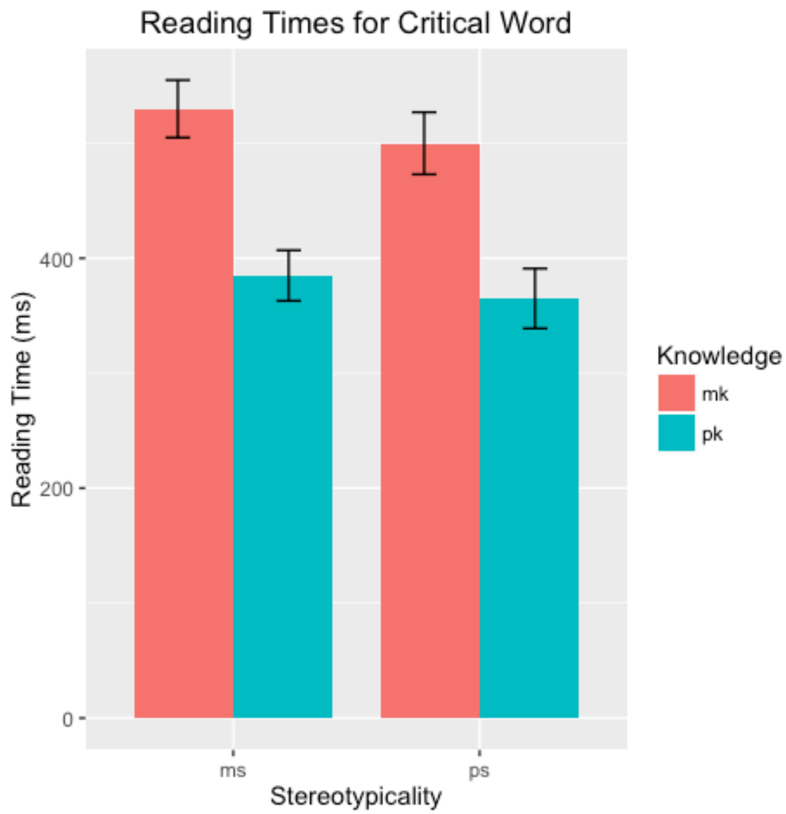
One main effect

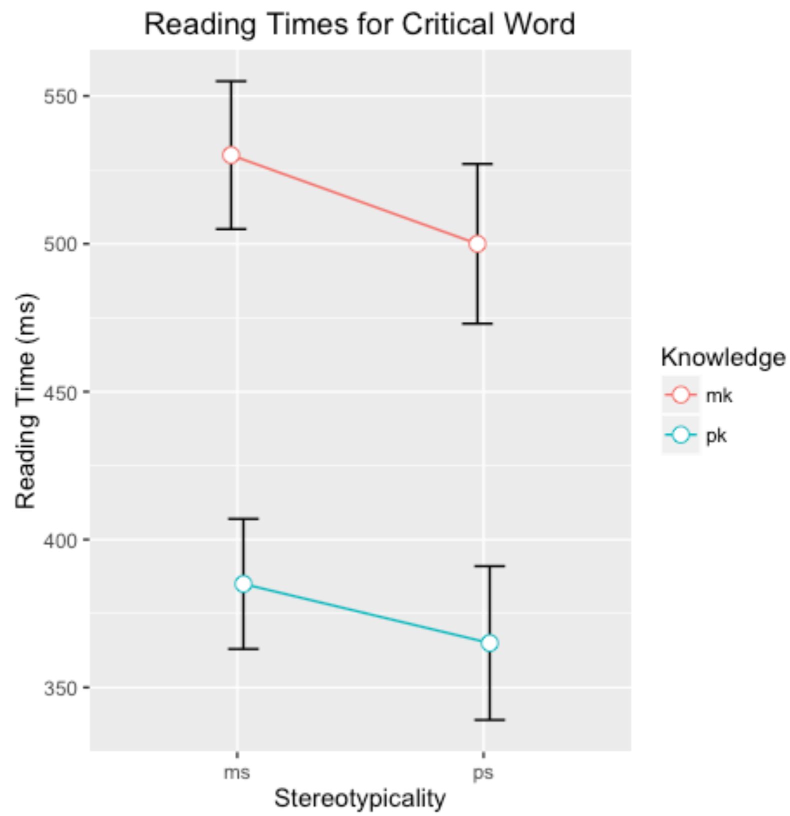




Possible Outcomes of a 2x2 Design

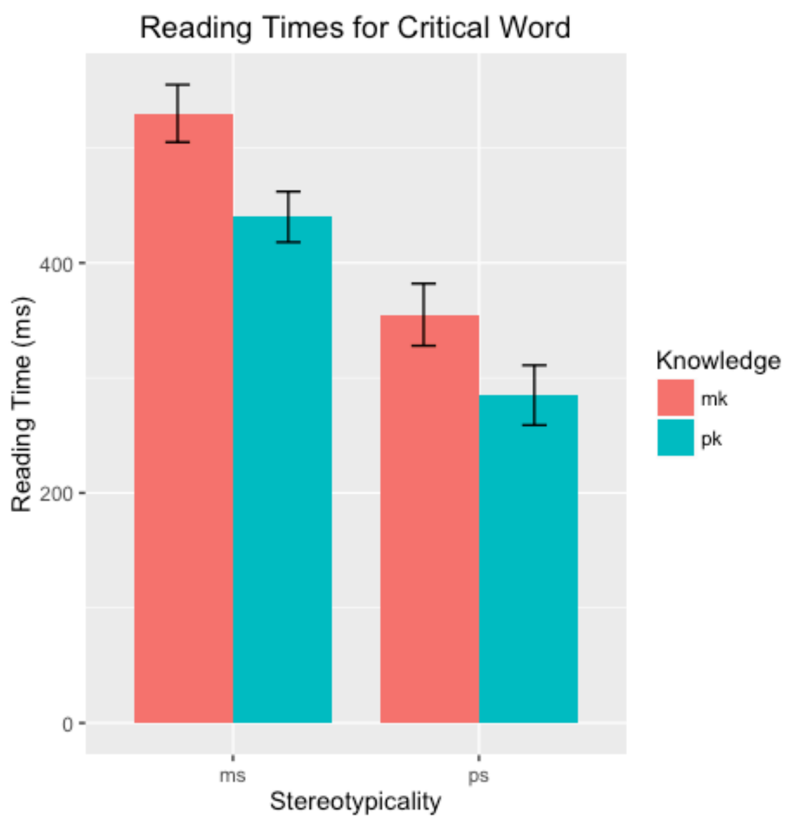
One main effect

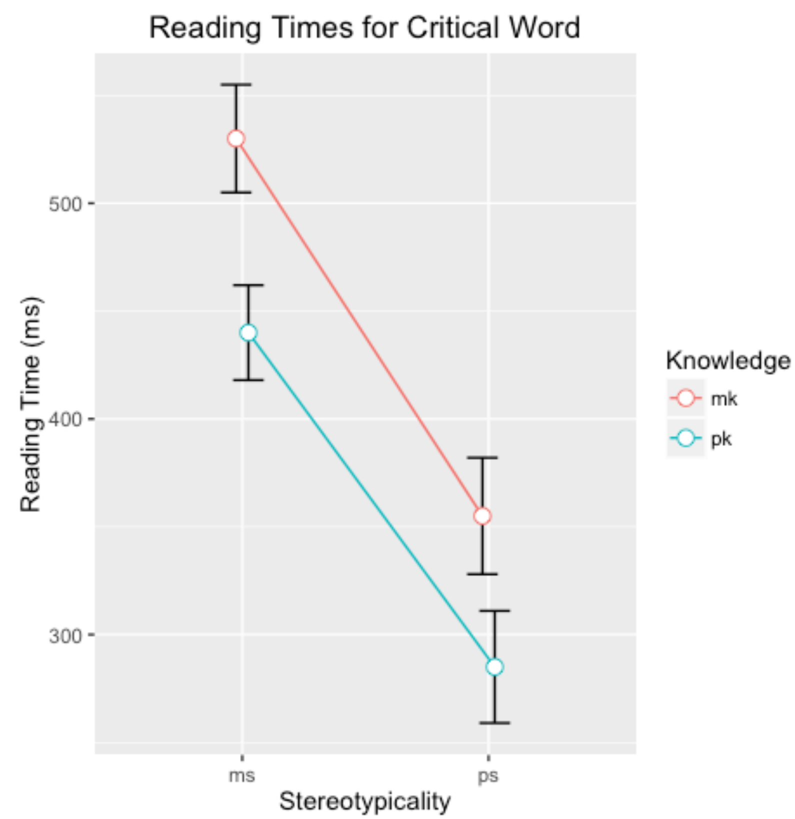




Possible Outcomes of a 2x2 Design

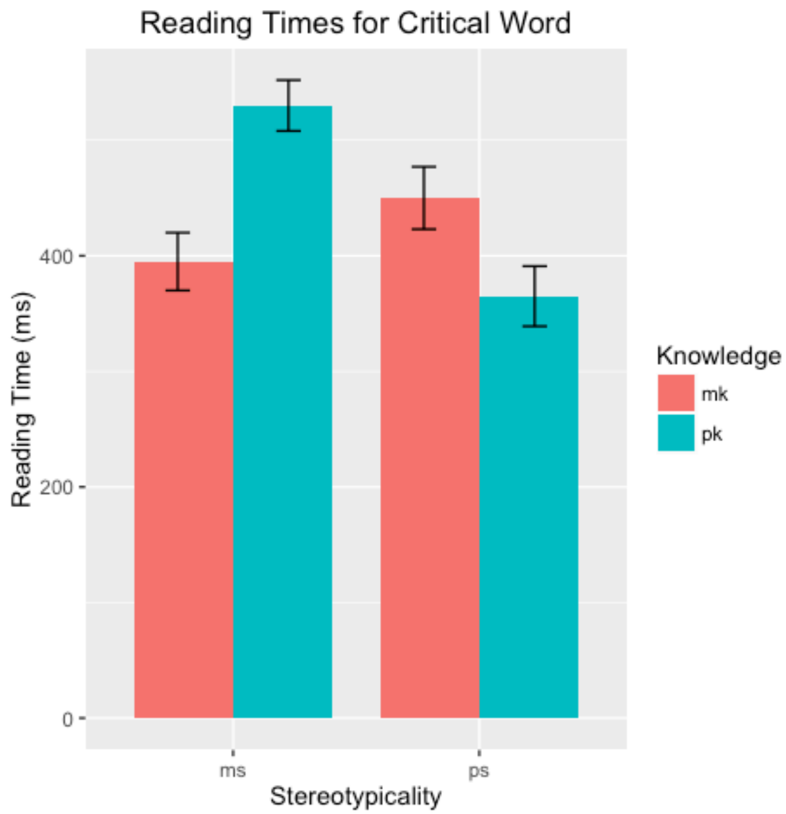
Two main effects

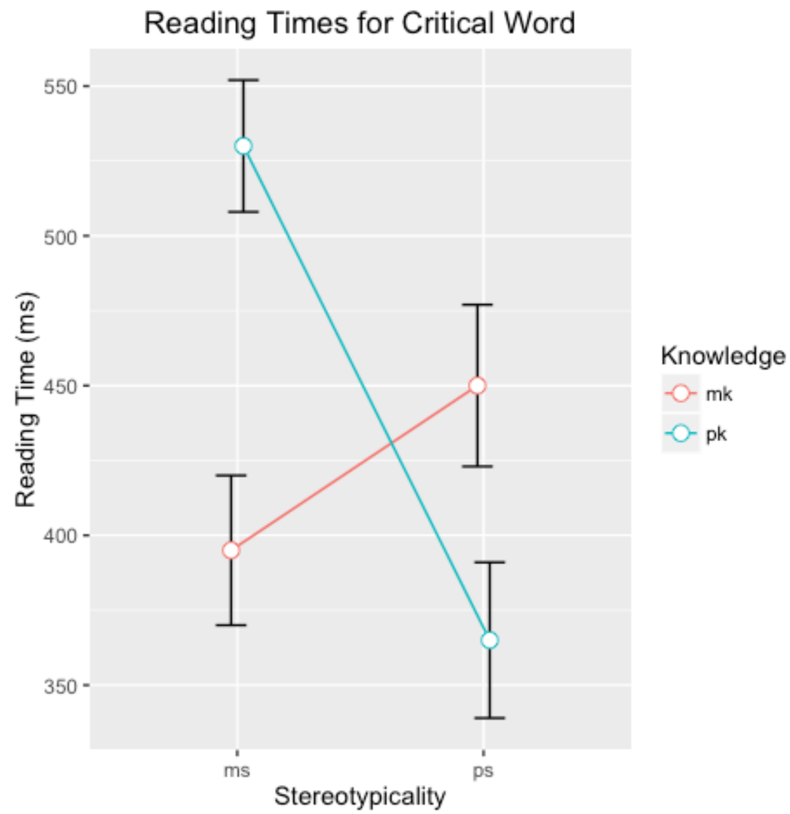




Possible Outcomes of a 2x2 Design

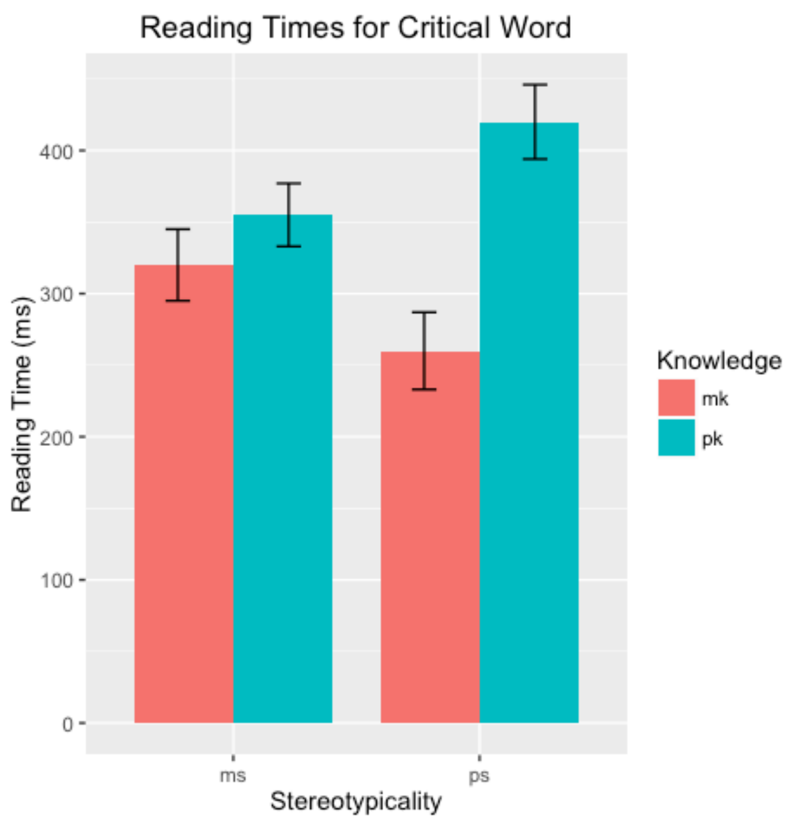
Interaction (but no main effects)

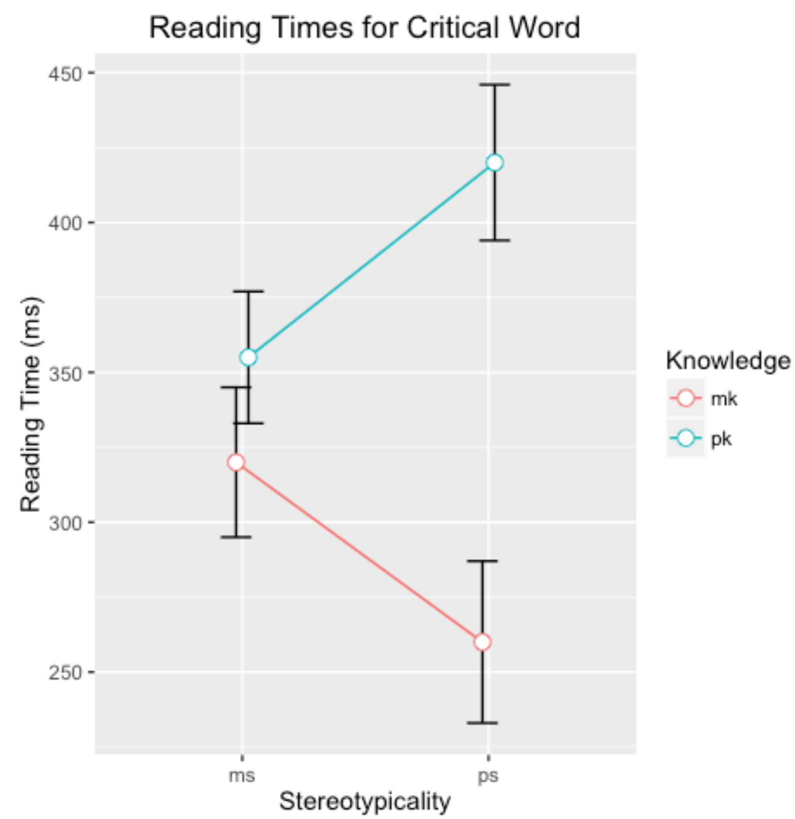




Possible Outcomes of a 2x2 Design

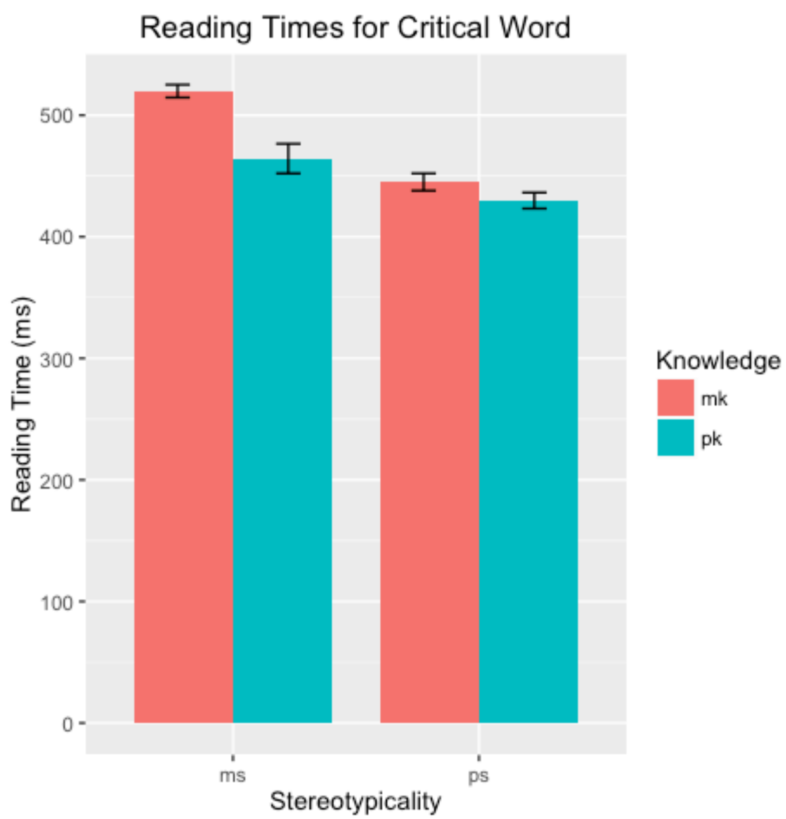
Interaction (but no main effects)

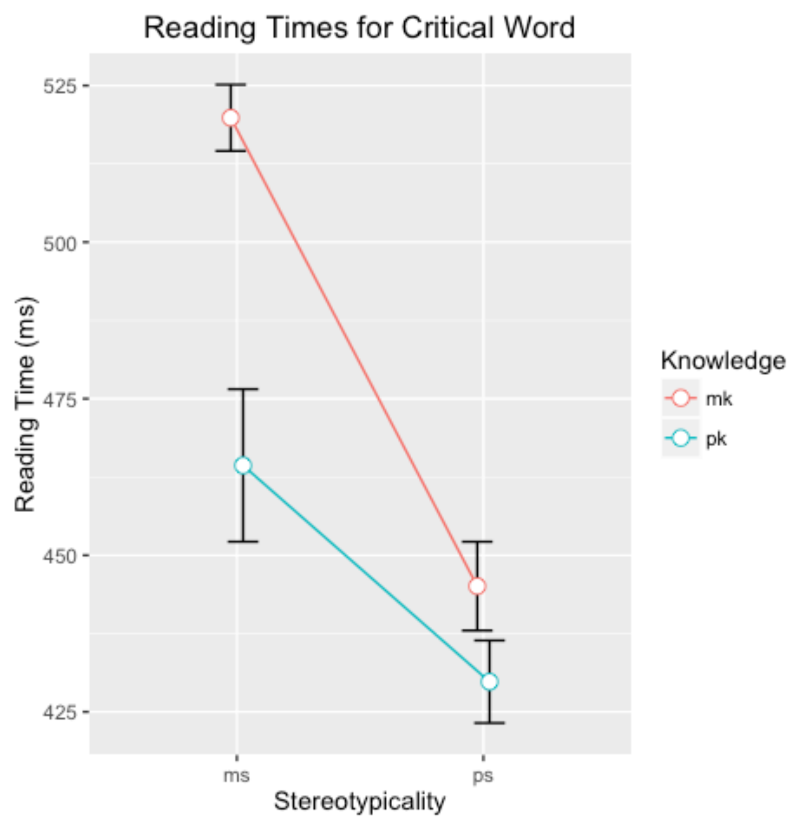




Plotting the results (hypothetical data)

Two main effects and an interaction





Interpreting the Results

Don't forget!

- **Graphs** – useful for identifying patterns and summarizing results
- **Statistical analyses** – tell us whether the results are likely to be "real"

Summary

Testing Hypotheses

1. State the null (H_0) and alternative hypotheses (H_1)
2. Set the decision criteria (e.g., choose alpha level)
3. Sample from a population (collect data)
4. Describe data and calculate appropriate test statistics
5. Make a decision (reject H_0 or fail to reject H_0)