Simple hypothesis tests

Two-sample t-test

Comparing sample differences

-

- If you find this value to be large, then the difference between samples is due to the experimental manipulation
- If you find this value to be small, then the difference is likely due to chance

The t-test for comparing two means

- Calculate the difference sample means $\bar{X}_1 - \bar{X}_2$
- Calculate squared standard deviations $S^2_{X_1}$ $S^2_{X_2}$
- Calculate combined standard deviation

$$S_{X_1X_2} = \sqrt{\frac{S^2_{X_1} + S^2_{X_2}}{2}}.$$

- Form ratio:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2} \cdot \sqrt{\frac{2}{n}}}$$

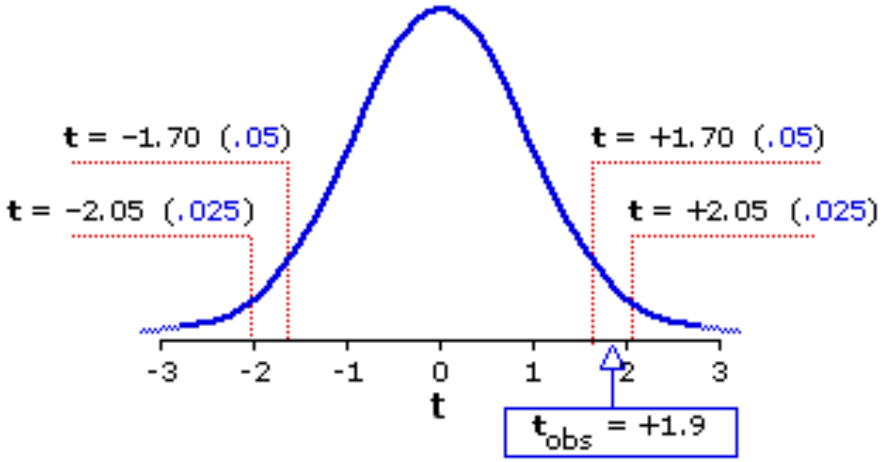
And here is sqrt(n) again!

- Look into t-distribution to find corresponding p-value for significance

Let's do the full thing

- We test the grade differences between two groups of students
- There are 28 students in each group → n=28
- The difference between the average grade in each group is 10 grade points
- The standard deviation in each group is 20 grade points
- So, we get t=1.9
- Now plot the t-distribution for n=28 and look for the probability that t is smaller than t=1.9

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1 X_2} \cdot \sqrt{\frac{2}{n}}}$$



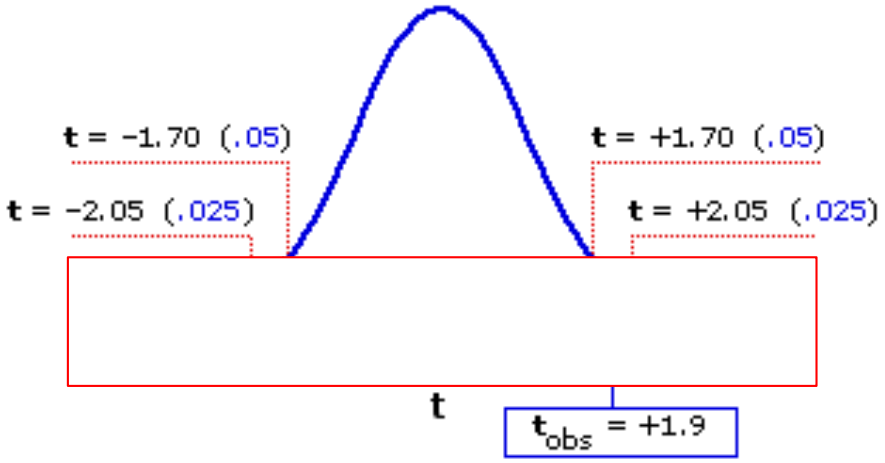
Level of Significance for a Directional Test				
.05	.025	.01	.005	.0005

df = 28	1.70	2.05	2.47	2.76	3.67

Directional versus non-directional

- Here's a slight, but important add-in – in which direction should the differences be??
- If we assume that one group must be **HIGHER** (or **LOWER**) than the other, we will be asking, what is the probability for which t<=1.9 (or t<=-1.9)!!
- **We can see that the cut-off for 5% is t=1.7 (or t=-1.7) in this case – and that means our groups are different!**

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2} \cdot \sqrt{\frac{2}{n}}}$$

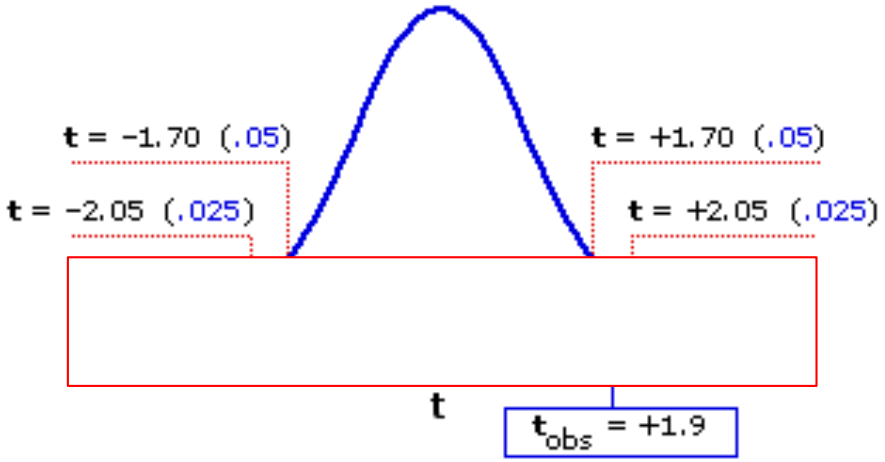


Level of Significance for a Directional Test				
.05	.025	.01	.005	.0005
df = 28	1.70	2.05	2.47	2.76	3.67

Directional versus non-directional

- Here's a slight, but important add-in:
- If we look for **DIFFERENCES** between the two groups, then the one group could be **HIGHER** or it could be **LOWER**
- So we are actually asking, what is the probability for which **|t|<=1.9!!!**
- **In this case, the cut-off is |t|=2.05, and our groups would NOT be different!!!**

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2} \cdot \sqrt{\frac{2}{n}}}$$



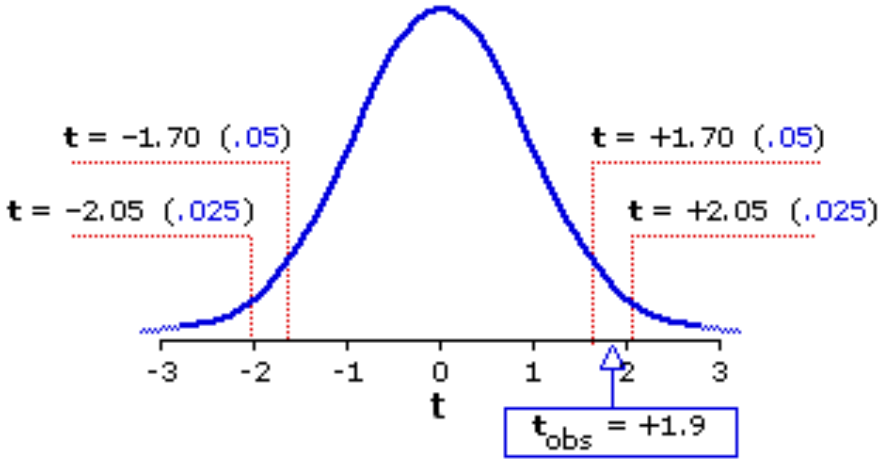
Level of Significance for a Directional Test				
.05	.025	.01	.005	.0005
df = 28	1.70	2.05	2.47	2.76	3.67

Directional versus non-directional

-

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2} \cdot \sqrt{\frac{2}{n}}}$$

- If we do NOT have a strong prior about the direction of the differences, we have to use the non-directional, or two-tailed test
- **If and ONLY IF** we have a STRONG prior about the direction of the differences, we are allowed to use the directional, or one-tailed test

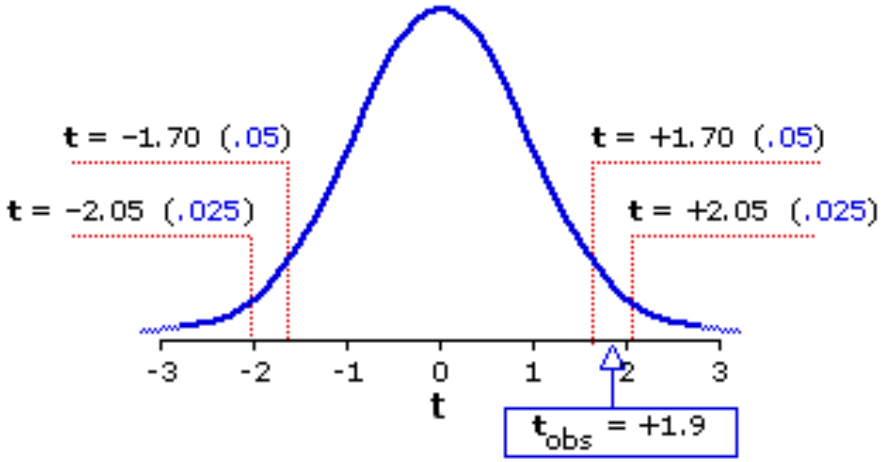


	Level of Significance for a Directional Test				
	.05	.025	.01	.005	.0005
	Level of Significance for a Non-Directional Test				
	---	.05	.02	.01	.001
df = 28	1.70	2.05	2.47	2.76	3.67

Decision

- For our data and the cut-off for an alpha=0.05, we have
- |t|=2.05 for the non-directional, two-tailed test
  - no significant differences in grades between the two student groups
- t=1.9 for a directional, one-tailed test
  - group A has significantly better grades than group B

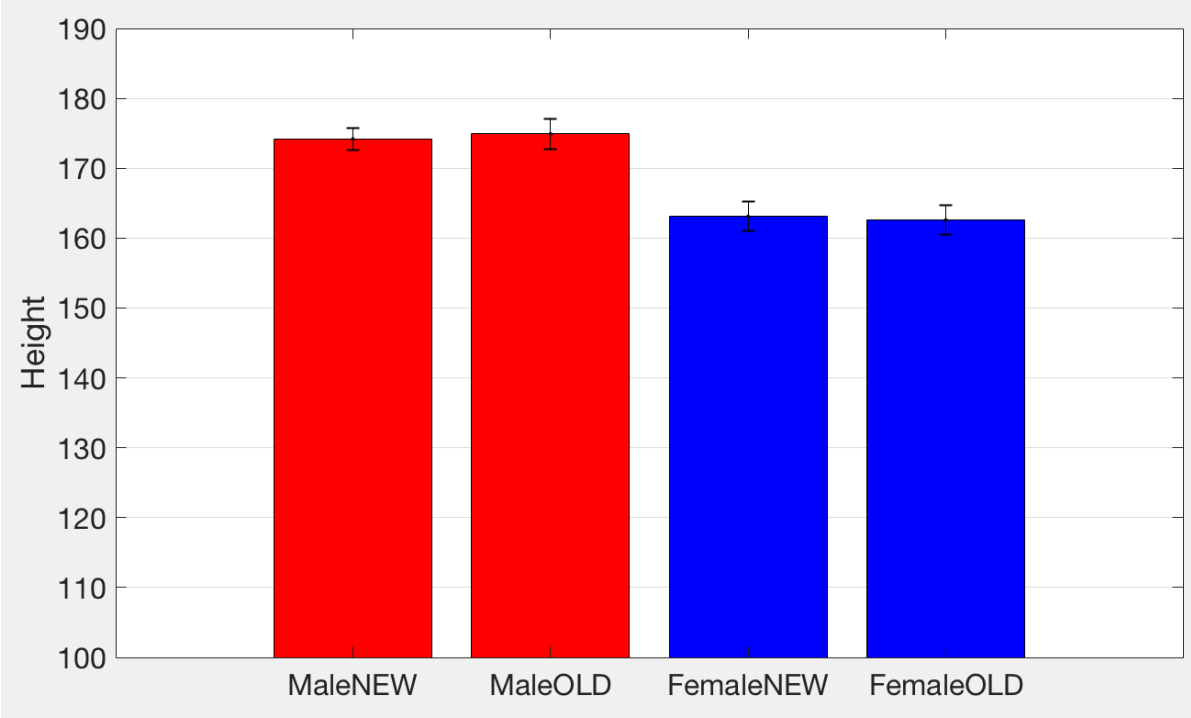
$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1 X_2} \cdot \sqrt{\frac{2}{n}}}$$



	Level of Significance for a Directional Test				
	.05	.025	.01	.005	.0005
	Level of Significance for a Non-Directional Test				
	---	.05	.02	.01	.001
df = 28	1.70	2.05	2.47	2.76	3.67

Our data

- For our data, we would like to know whether male and female heights are different
- Using the two-sample t-test, we find for the data from this class
  - t_{56}=8.607, p<<.001
- for the data from last class
  - t_{49}=8.143, p<<.001
- Hence, we conclude that the average heights are different
- Note, we could have asked whether height is larger or smaller
  - one-tailed test, simply divide p by 2



Assumptions of two-sample t-test

- Samples were drawn independently
  - t-test for paired samples available
- Variance in data is the same in both groups
  - use t-test for unequal variances
- Sample sizes are the same
  - corrected t-test for unequal sample sizes
- **Data follow a normal distribution**
  - **use non-parametric test otherwise**

Effect size

More about the meaning of life

- Classical example from Cohen's excellent paper "Things I've learned (So Far)". American Psychologist (45)12, 1304-1312, 1990.
- "Children's Height Linked to Test Scores": 14,000 children (6 to 17 years of age) tested: found *significant* link between height (age- and sex-adjusted) and scores on tests of both intelligence and achievement (adjusted for socioeconomic status, birth order, family size, and physical maturity)

More about the meaning of life

- If significant means $p<0.001$, then $r=0.0278$ is highly significant for the 14000 children that they measured
- Turning correlation into causation:
  - increasing IQ from 100 to 130 means growing height by 5 meters
  - increasing height by 10 centimeters: need to raise IQ by 900 points
- Obviously, this is silly!
- Significance (p-value) is not the most important thing for publication!!

Effect Size

- Effect size refers to the **practical**, rather than the statistical, significance of the relationship between variables
  - Is the effect “real”?
  - Does it affect models of perception and the brain?
  - Can it change the way we think about things and how we act?
- However, the effect size, like any other statistical measure, varies from sample to sample
  - Repeating one study 10 times yields 10 different effect sizes
- Be cautious in interpreting effect sizes as well!!

Calculating effect size

- Different statistical tests have different effect sizes developed for them
- The general principle is always the same

Effect size is the magnitude of the impact of the independent variable (factor) on the dependent variable (measurement)

Effect size for comparing two means

CognitiveSystems

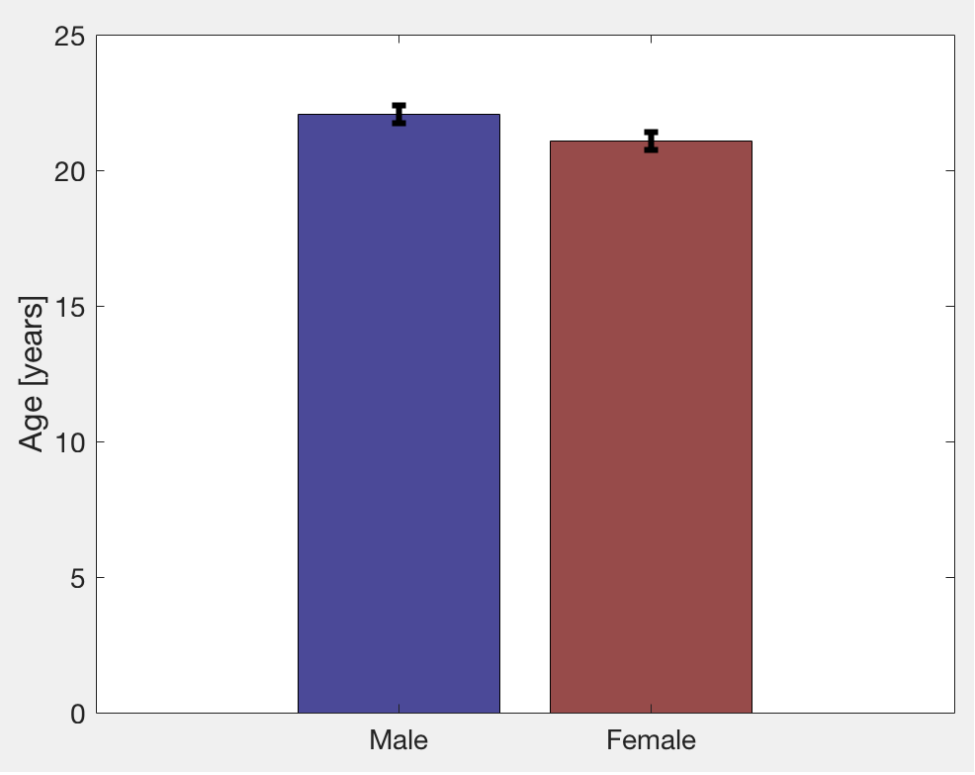
- *d*: Focused on standardized mean differences
  - Allows comparison across samples and variables with differing variance
- For our data **on heights**, the effect size is very large (since the difference between the two means is large and we have large sample sizes)!

$$d = t * \sqrt{\frac{1}{n} + \frac{1}{m}} = 8.607 * \sqrt{\frac{1}{34} + \frac{1}{24}} = 2.29$$

Effect size	d
Small	0.3
Medium	0.5
Large	0.8

Effect size for comparing two means

- Let's compare the means for the **Age distributions** for males and females in our data
  - m(male)=22.06, m(female)=21.08
  - t(56)=2.031, p=.047
  - males are differently old in this class!



Effect size for comparing two means

- The effect size for this is medium!, but compared to the height difference, our age effect is much smaller!

$$d = t * \sqrt{\frac{1}{n} + \frac{1}{m}} = 2.031 * \sqrt{\frac{1}{34} + \frac{1}{24}} = 0.54$$

Effect size	d
Small	0.3
Medium	0.5
Large	0.8

Key concepts

- Whenever you interpret your results, make sure to look at the size of the effect
- Effect size values are “normalized” and hence allow you
- Never stop with the p-value and significance, but critically question how large the effect really is and what impact this has for your scientific question!!

Simple hypothesis tests

Analysis-of-Variance (ANOVA)

ANOVA (ANalysis Of VAriance)

- Extension of t-test
- Compares differences **between more than two groups** (or samples)
  - independent variable (also called “factor”) has more than three “levels”
  - dependent variable is measured (often also called “response variable”)
- Analysis of choice for factorial designs
  - remember our Experimental Design lecture?

What does ANOVA do?

CognitiveSystems

- Splits variability in measured data into:
  - factor-related variability (due to differences in the groups)
  - “unexplained” variability (due to random differences between observations within a factor)
- Example: Experiment in which we compare the perceived trustworthiness of male faces due to five hair colors (brown, black, red, blond, yellow)
  - Variability in trustworthiness due to hair color
  - Variability due to other factors (individual differences within one hair color)

One-factor ANOVA

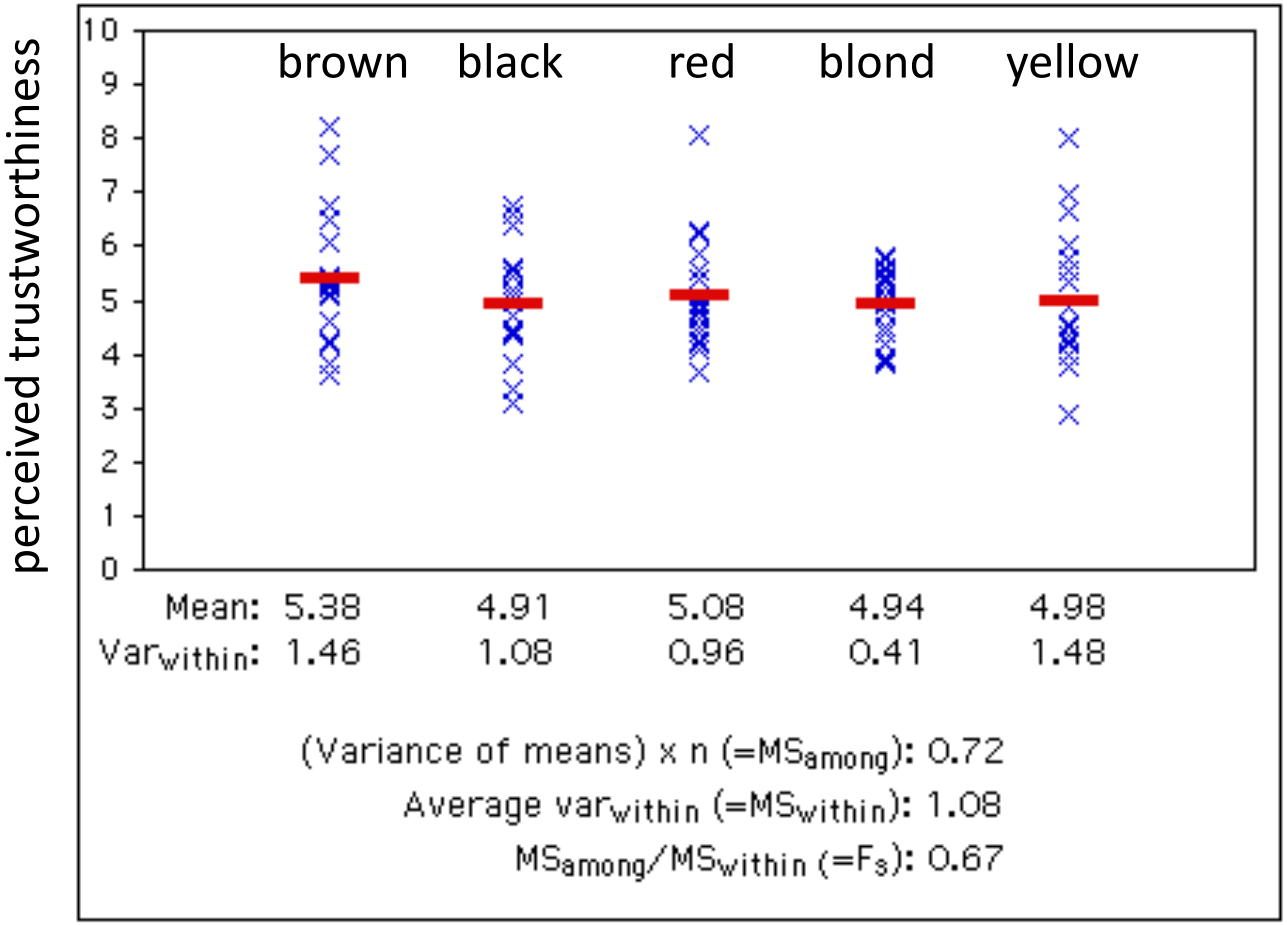
- One independent variable with more than two levels
- H_0: trustworthiness does **not** depend on hair color
- Form so-called "F-ratio":

$$\frac{\text{variability between hair colors}}{\text{variability within hair colors}} = \frac{\text{variability due to hair}}{\text{variability due to randomness}}$$

- If F-ratio is [ ] hair color likely has an influence on trustworthiness
- If F-ratio is [ ] trustworthiness most likely does not depend on hair color

One-factor ANOVA

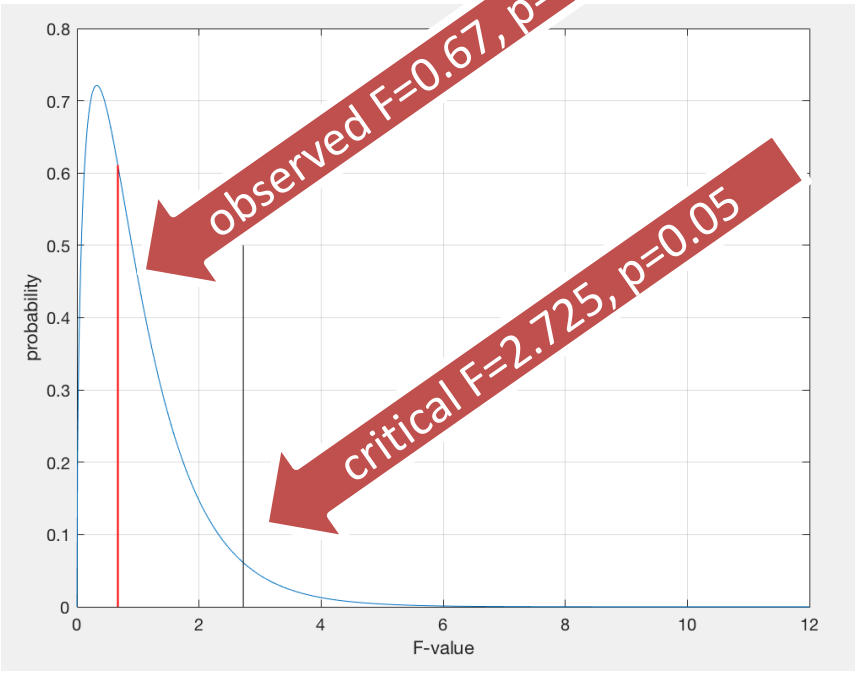
- Five samples (n=20) from populations with parametric means of 5. Red bars indicate sample means.



http://udel.edu/~mcdonald/statpix/statanovasig1.gif

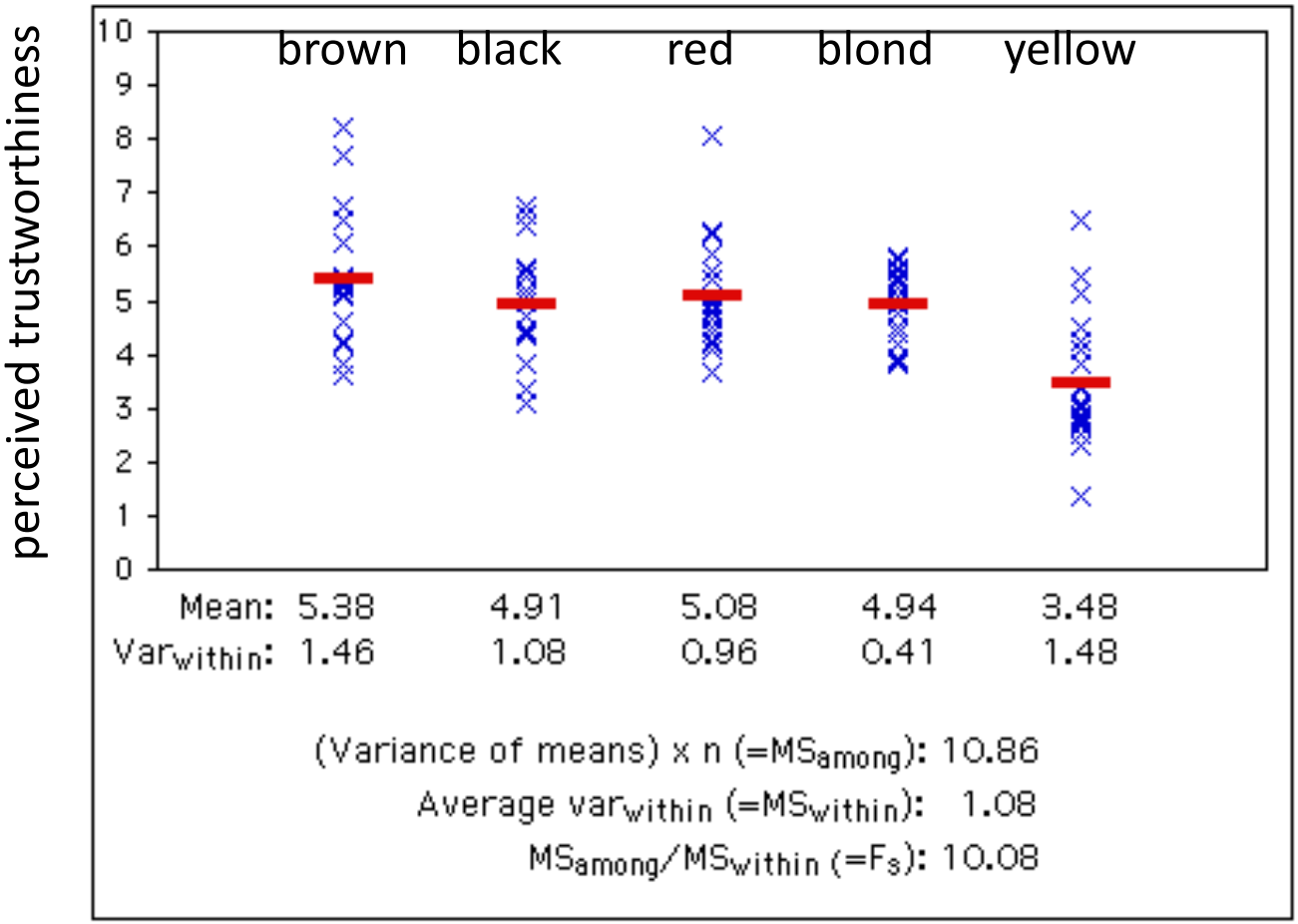
One-factor ANOVA

- We look for MS_{among}/MS_{within} factor, which is the "F-ratio"
- We now look up this value in an "F-distribution" and repeat the exercise we've done many times, looking for the probability of obtaining such a value "by chance"
- The result is not at all significant, as we can see in the plot (the p-value is 0.5731)
- So, we conclude that hair-color in this experiment does not have an effect on perceived trustworthiness!



One-factor ANOVA

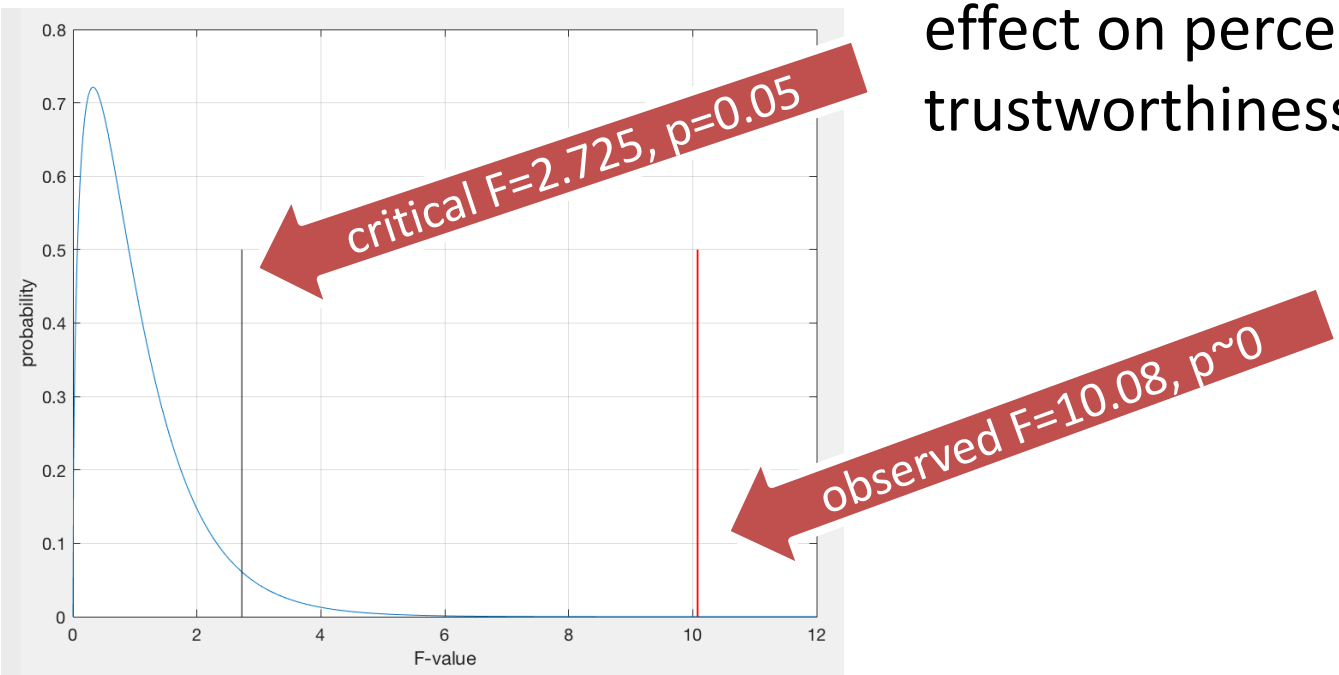
- Four samples (n=20) from populations with means of 5; the last from one with a mean of 3.5. Red bars are sample means.



http://udel.edu/~mcdonald/statpix/statanovasig2.gif

One-factor ANOVA

- Compared to the first sample, we only shifted one group's data downwards, and we saw that this has a large effect on the MS_{among}/MS_{within} factor
- This time, the result is highly significant, as we can see in the plot (the p-value is virtually 0)
- So, we conclude that hair-color in this experiment would have an effect on perceived trustworthiness!



Two-factor ANOVA

- 2 independent variables / factors with more than 1 level for each factor
- The two-factor ANOVA partitions variability:
  - within a level of a factor
  - between levels of a factor
  - between levels of a factor and between levels of the other factors –
- By relating how large the variability is in the effects, one can also calculate the effect sizes for each effect
  - this is usually called η^2

Two-factor ANOVA

- Experiment: trustworthiness ratings depending on gender (M & F) and eye size (Large & Small)

- Variability betw M & F (effect of Factor 1 “gender”)
- Variability betw L & S (effect of Factor 2 “eye size”)
- Variability betw M & F for L vs S, or equivalently
- Variability betw M vs F for L & S (interaction effect)

- What is the effect of each factor on ratings?
- If there is an effect, is it the same for all levels of the other factor?

2 x 2 factorial design - Actual Numbers

	Gender (A)		
Eye Size(B)	**Male (A_1)**	**Female (A_2)**	**Row Means (effect)**
Small (B_1)	3.3	2.5	2.9
Large (B_2)	5.3	6.1	5.7
Column means (effect)	4.3	4.3	4.0

Main Effects

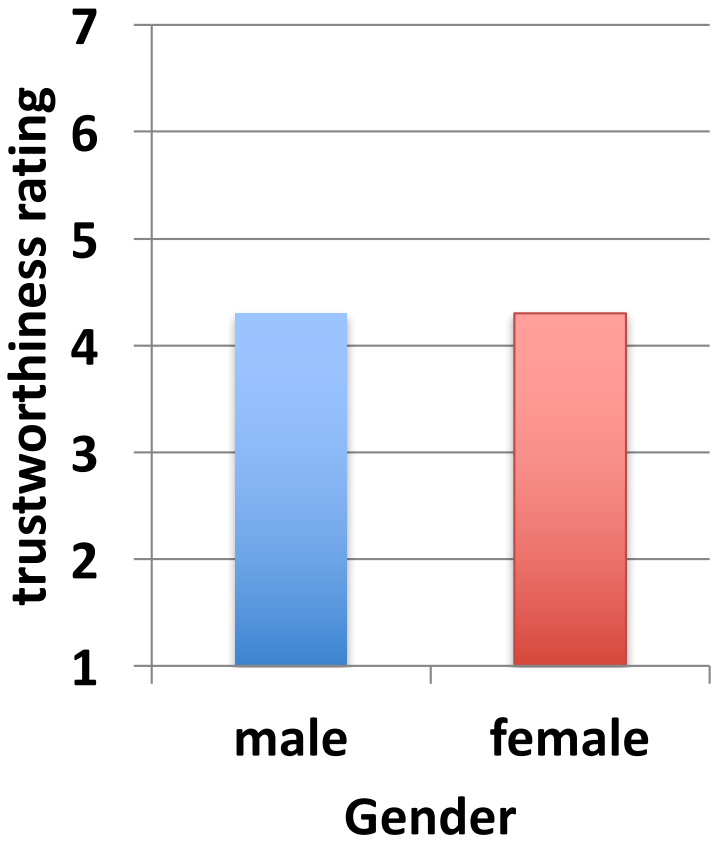
- Main Effect
  - The effect of one independent variable averaged over all levels of another independent variable
- In the examples, these would be:
  - Column (Gender) main effects
  - Row (Eye Size) main effects

Interactions

- Interaction
  - Happens, whenever the effect of one independent variable depends on the level of another independent variable
  - Often shown using graphs
- If you find an interaction, the main effects cannot be interpreted without discussing the interaction as well!!

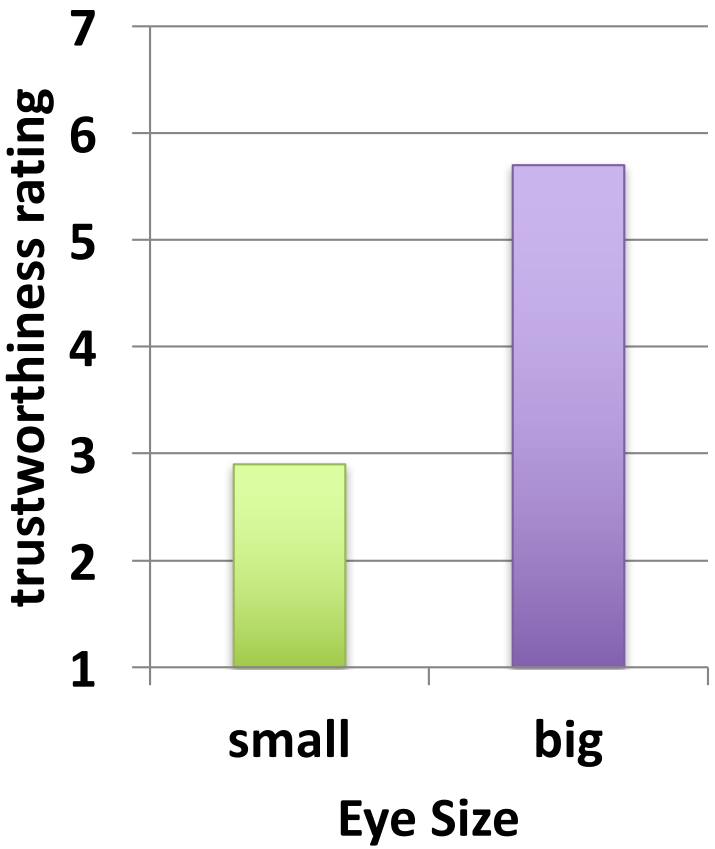
Column Main Effects

- The dependent variable is on the Y-axis
- The column main effects are obtained by averaging over the row in a 2-way design
- It seems that there is **no effect** of gender on attractiveness ratings



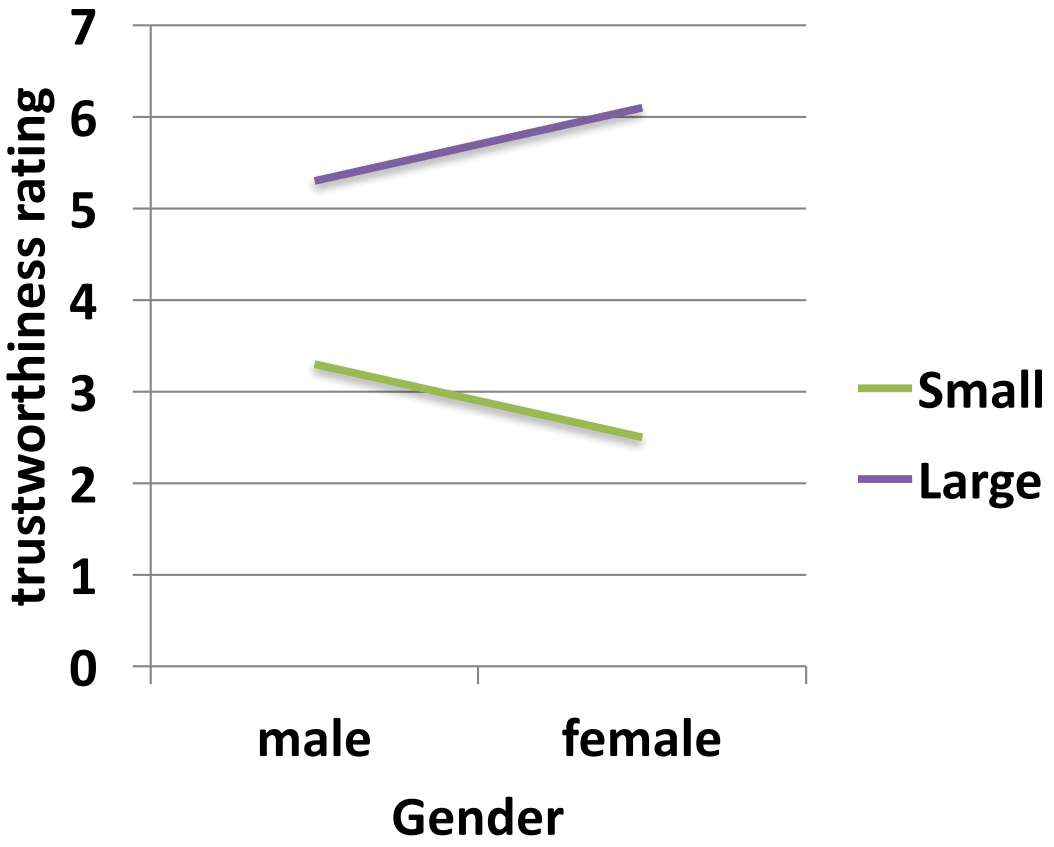
Row Main effects

- The row main effects are obtained by averaging over the column in a 2-way design
- Small eyes are less trustworthy than large eyes
- This needs to be tested for statistical significance (see later)



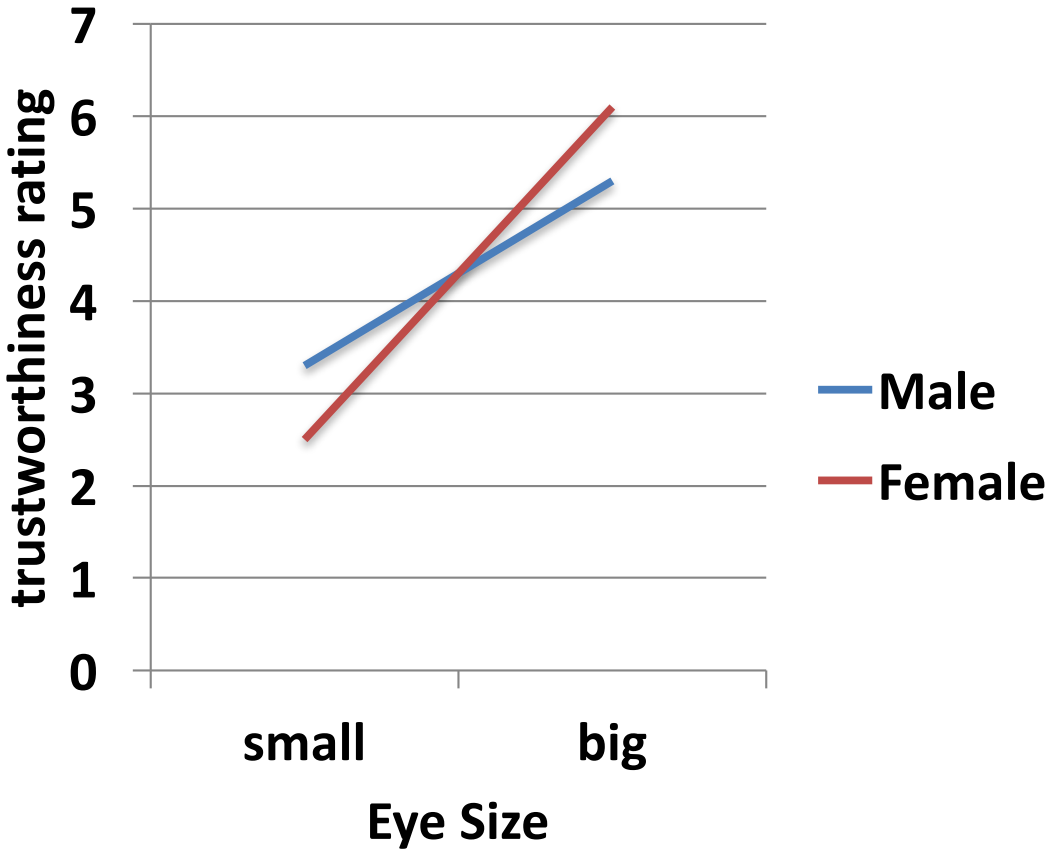
Interaction of Rows and columns

- If the graph shows lines that **are not parallel** then there is an interaction
- The effect of eye size is much larger in females than males



Interaction of Rows and columns

- If the graph shows lines that **are not parallel** then there is an interaction
- The effect of eye size is much larger in females than males



Interpretation

- If you do the actual statistics, you would find that:
  - no main effect of gender
  - a significant main effect of eye
  - the main effects are qualified by a significant interaction between eye size and gender
- **In analyzing the results, we cannot say that gender has no influence on trustworthiness, but we must**

Interpreting the ANOVA result

- Reporting the statistics is only a part of the story
- The most important task for the scientist is to interpret the meaning of the statistical results and to set them into the scientific context
- This requires
  - a thorough understanding of the limitations of the design and the statistical tests
  - a good grasp of the relevant background literature
  - an idea of how important this result is

Interpreting the ANOVA result

- In the present case, the study investigated trustworthiness rating of male and female faces varying in eye size
- Questions related to the data/design:
  - how was trustworthiness defined for participants?
  - can we treat the ratings as normally distributed numeric data?
  - what was the population studied (e.g., WEIRD)?
  - is the stimulus set representative? (of the population? of all faces?)
  - how well did people agree in their trustworthiness ratings?
  - was this a fast or slow decision (look at response times!)?

Post-hoc tests and multiple comparisons

Post-hoc analyses

- Let's go back to our one-way ANOVA with hair color – let's say we have a significant effect – but now what? Since we have five levels of hair color, we do not know WHICH color may have made the difference!
  - remember: an ANOVA just says: "somewhere there's a difference"
- This means, you need to run **post-hoc tests**
- Post-hoc tests are different from a-priori tests, in which you specify BEFORE the experiment, which conditions will be expected to be different
- This is a subtle, but important difference
  - a-priori tests do NOT need to be corrected for multiple comparisons
  - post-hoc tests need to be corrected for multiple comparisons

Example

- You are the CEO of a major pharmaceutical company and you want to develop a drug that helps peoples' concentration – your drug designers go to work and come back reporting that their new drug does not significantly improve concentration, at the usual alpha level of .05
- "Argh", you think, and instruct the designers to look for effects of this expensively-developed drug for 500 more kinds of effects, all at alpha=.05
- The designers come back and report that, lo and behold, they found 26 effects in which the drug worked!
-

Multiple comparisons

- If you do 100 statistical tests, for which H_0 is actually true (no effect in the real world), you will find ~5 tests for which $p<0.05$ just by chance (you report a false alarm)!
- If you have 10 conditions, for which you would like to compare every condition to every other condition, these are already 90 t-tests!!
- Solution: correct for multiple comparisons

Multiple comparisons

- The easiest way is to simply adjust your alpha level, that is, to control the chance of getting a false positive
- The most conservative way of doing this is the
  - if you do N tests, simply adjust your alpha to alpha / N
  - for 10 tests, any test with **p < .005** would then be significant
  - for many tests, this will lead to very low values, but almost guarantees a low level of false positives
- There are other, less conservative ways of adjusting

A recent case

- Analysis of percentage of participants who cheated following “priming” by money-related, neutral, or time-related words.
- They found that “The percentage of participants who cheated varied across conditions, $\chi 2(2, N = 98) = 14.61$, $p = .001$ (see Fig. 1)”

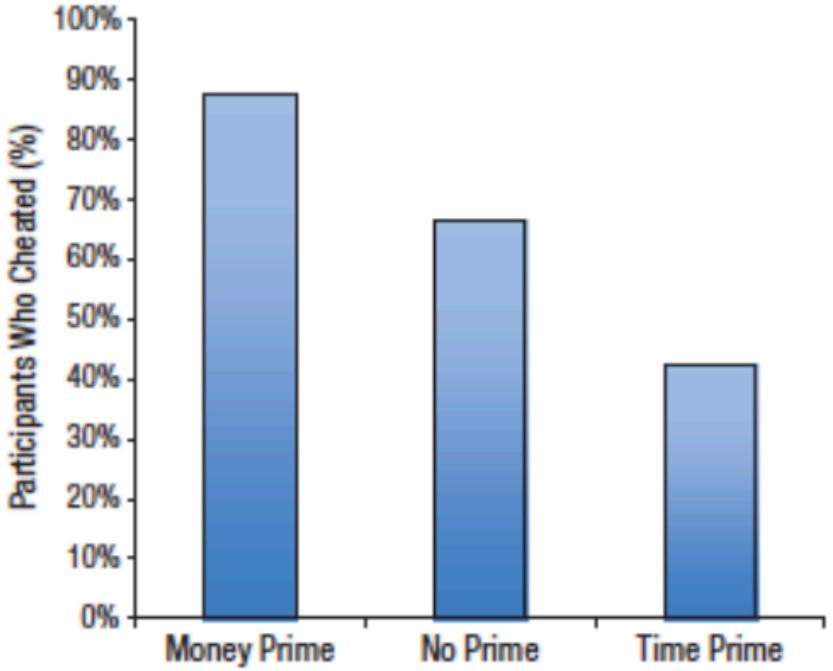


Fig. 1. Percentage of participants who cheated in each condition of Experiment 1.

Gino, F. & Mogilner, C. (2013). Time, Money, and Morality. Psychological Science (25) 2, 414-421.

A recent case

- Analysis of percentage of participants who cheated following “priming” by money-related, neutral, or time-related words

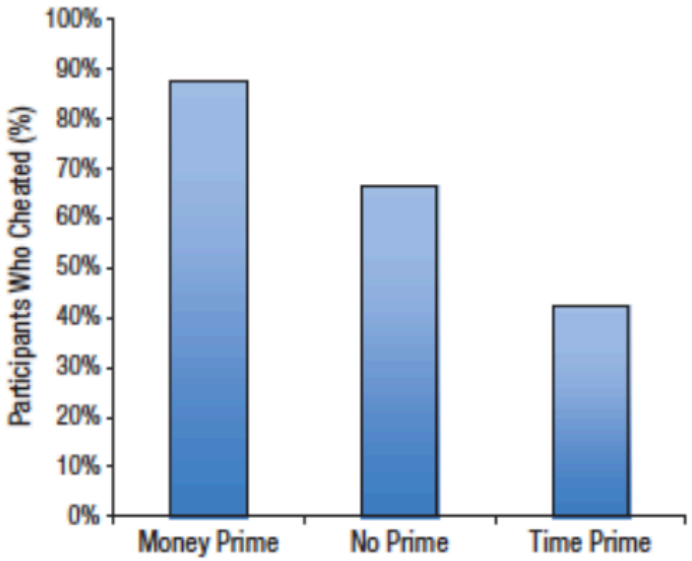
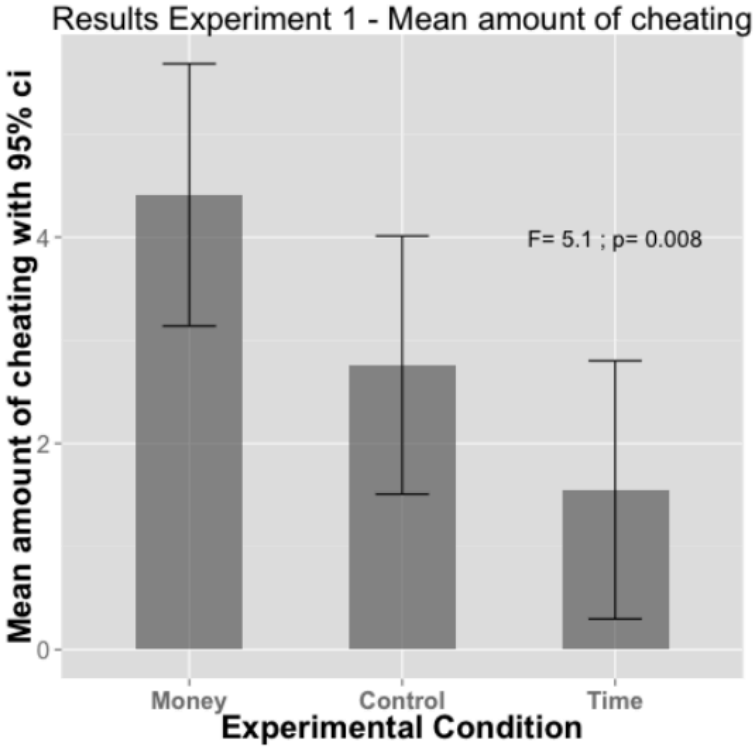


Fig. 1. Percentage of participants who cheated in each condition of Experiment 1.

No confidence intervals!

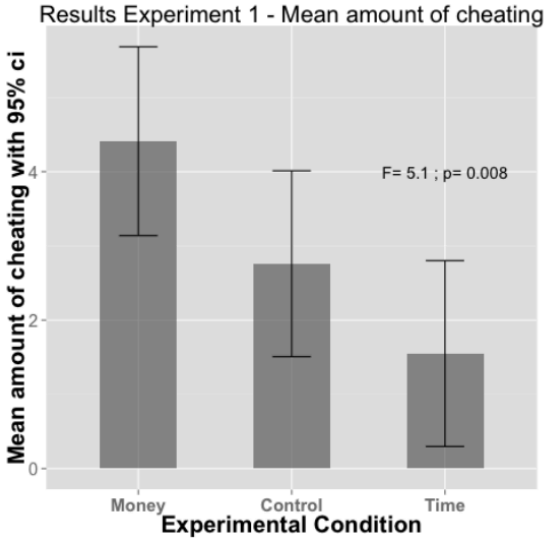


Proper Figure re-drawn

Gino, F. & Mogilner, C. (2013). Time, Money, and Morality. Psychological Science (25) 2, 414-421.

A recent case

- You need to compare money to control and time condition, as well as time to control condition = 3 tests
  - best to divide alpha (0.05) by 3 (Bonferroni)
- In the paper, however, they report tests that were made WITHOUT correcting for multiple comparisons!
- If you redo the tests, none of the two experimental conditions will be different from the control condition!



Gino, F. & Mogilner, C. (2013). Time, Money, and Morality. Psychological Science (25) 2, 414-421.
http://people.stat.sfu.ca/~cschwarz/CourseNotes/Reanalysis/TimeIsNotMoney/time-is-money-error.pdf

Keypoint

- Every time you run multiple tests involving comparisons of the same dataset, you should correct for multiple comparisons
  - This is especially important, for example, in imaging studies, where you run tests on millions of voxels in the brain!
- **As usually alpha is affected by the correction, correcting for multiple comparisons greatly reduces**

Non-parametric tests

Non-parametric tests

- Used for ordinal data (such as starred movie reviews, assessing preferences, etc.)
- They make fewer assumptions about the underlying probability distribution, and therefore are more robust.
- This comes at a cost, however: in cases where a parametric test would be appropriate, non-parametric tests have **less power**.
  - A larger sample size is needed to draw conclusions with the same degree of confidence.

Non-parametric tests

- Each of the above-mentioned parametric tests (t-test and ANOVA) has its non-parametric counterpart
- Since both t-test and ANOVA assume normally-distributed population data, sometimes it can be advisable to test your sample data for normality
- If this is not possible (when you have a small sample, for example), or if your data is ordinal then you SHOULD ALWAYS use non-parametric tests

Example

- A teacher decides to test the prediction that parental guidance on drawing will influence neatness by **rank ordering** her 21 students according to the neatness and organization of their assignments. She consulted the parents in categorizing the students to the strong guidance group (A) and the less guidance group (B)
- This data is rank ordered and hence it is better to compare the two groups' average ranks using a non-parametric test

When to do what?

- If you believe raw data are not normally distributed, use a non-parametric test
  - this is usually safer – even for larger populations!
- If you believe your data are normally distributed, but you **don't know the standard deviation** of the population a-priori, use the t-test
- If you believe your data are normally distributed and you **know the standard deviation** a-priori, you can use the z-test that we looked into earlier
  - but, using the t-test will be very fine in any case!

Key concepts

- Simple hypothesis tests for comparing means
  - basic idea behind the test
  - how do I conduct a t-test? (no equations, just the logic)
- Effect size
  - why is it important?
  - what does it allow me to do?

Key concepts

CognitiveSystems

- ANOVA
  - one-factor ANOVA: basic idea of comparing variabilities
  - two-factor ANOVA: distinction between main effect and interaction
- Post-hoc tests and correction for multiple comparisons
  - why is it necessary?
  - what part of the test is affected by the correction?
- Non-parametric tests
  - when to choose and effects on power