Data Analysis

Basic Statistics & Excel

PSYC260 Lab / Winter 2010 / Melissa Ellamil

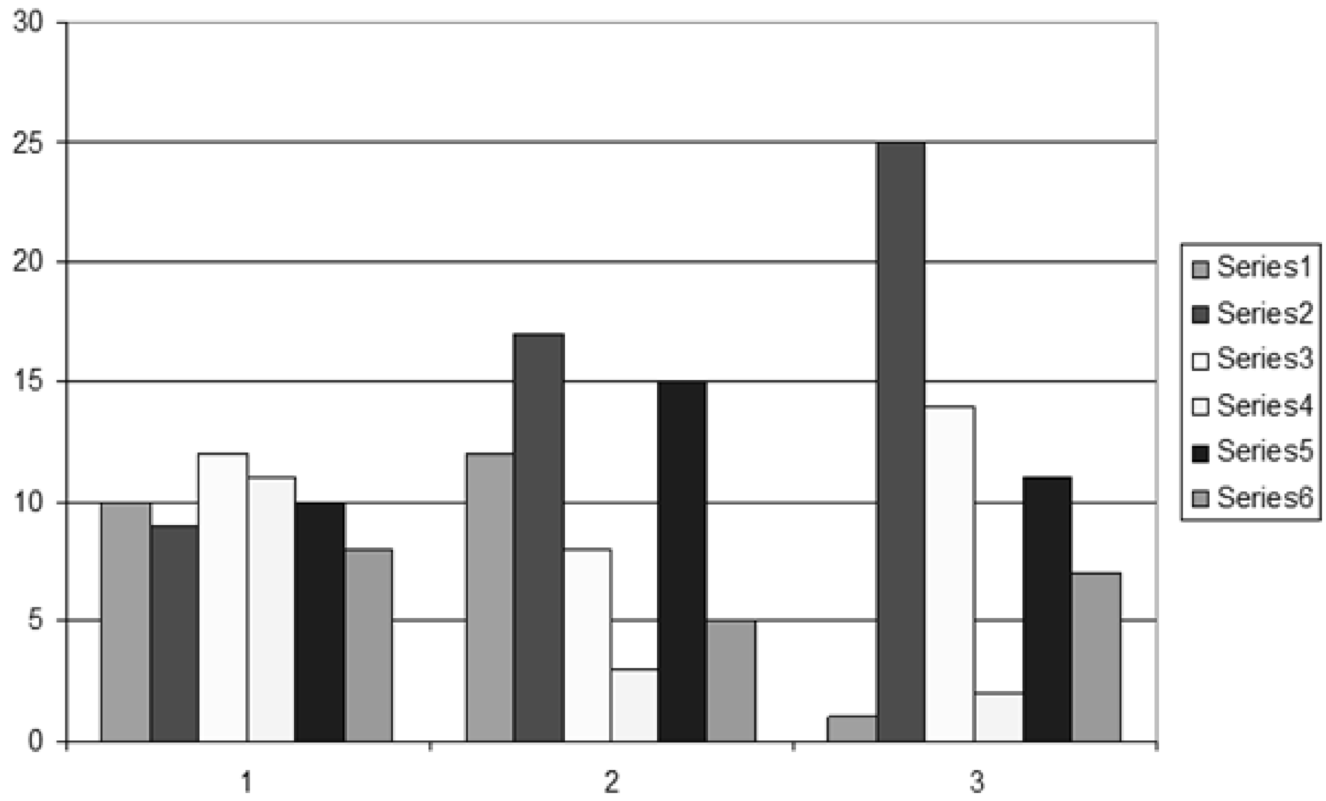
Descriptive Statistics

- Mean (Average)
  - Central value that the data tend to cluster around
  - Sum of all scores divided by the number of scores
- Standard deviation
  - How much variation is in the data
  - On average, how far the scores are from the mean
- Correlation
  - Size of relationship between two variables (0 to 1)
  - Type of relationship (positive or negative)

Mean & Standard Deviation

	Group 1	*Group 2*	*Group 3*
	10	12	25
	9	17	11
	12	8	1
	11	3	2
	10	15	14
	8	5	7
Mean	**10**	**10**	**10**
SD	**1.41**	**5.58**	**8.90**

Mean & Standard Deviation



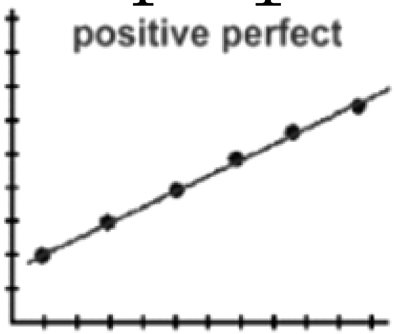
Mean & Standard Deviation

- Mean (Average)
  - Enter the numbers in columns in Excel
  - Click on the cell where you want the mean to be
  - Type `=average(first cell of column:last cell of column)`
    - *E.g.*, `=average(b2:b9)`
  - **OR** Highlight the cells of the column you want to include, instead of typing the cell range

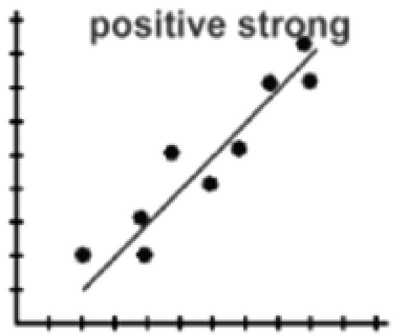
- Standard deviation
  - `=stdev(first cell:last cell)`

Correlation

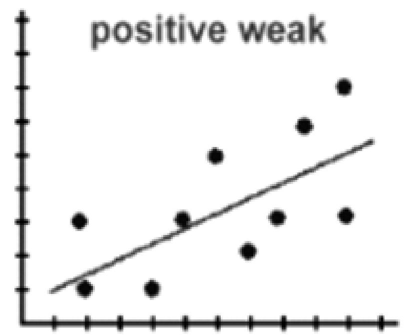
r = 1



This could show how the distance travelled in a vehicle increases as time increases, if the vehicle maintains a constant speed.

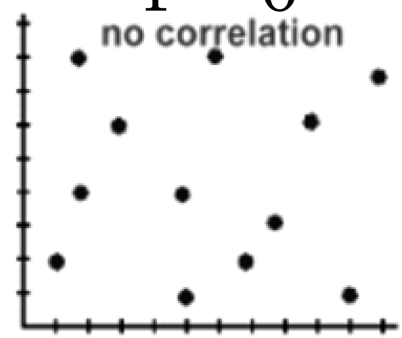


This could show the increase in a student's height as their grade level increases.



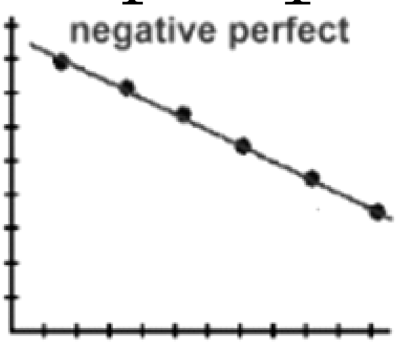
This could show the increase in a person's weight as age increases. Notice there is more variability here, as more factors can affect a person's weight.

r = 0

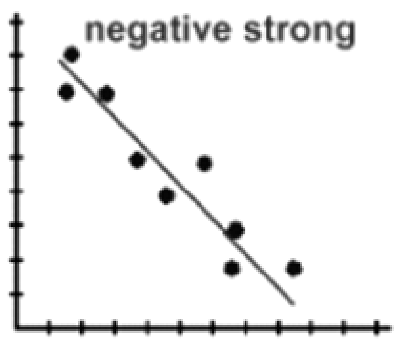


This could be the result if you compare a person's Math mark to the number of pets they have. There's no connection.

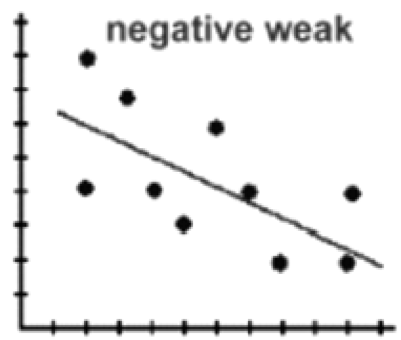
r = –1



This could show how the acceleration of an object decreases if the mass increases, if the force applied is constant.



This could show the decrease in a trucker's net income as the cost of gas increases.



This could show the decrease in the no. of vehicle accidents a person has, as their age increases past age 25.

CORRELATION

- Enter the numbers in columns in Excel
- Click on the cell where you want the correlation
- Type `=correl(first cell:last cell of column 1, first cell:last cell of column 2)`
  - *E.g.,* `=correl(`**`b2:b11,c2:c11`**`)`
- **OR** Highlight the cells of the column you want to include, instead of typing the cell range

Excel Data Analysis

- Excel 2007
  - Click the *Microsoft Office Button*, and then click *Excel Options* (at the bottom right corner)
  - Click *Add-ins* (on the left), and then in the *Manage* box (near the bottom), select *Excel Add-ins* and click *Go*
  - In the *Add-Ins Available* box, select the *Analysis ToolPak* check box, and then click *OK*.
  - If you are prompted that the *Analysis ToolPak* is not currently installed on your computer, click *Yes* to install it
  - When you load the *Analysis ToolPak*, the *Data Analysis* command is added to the *Data* tab

Excel Data Analysis

- Excel 2003
  - On the *Tools* menu, click *Add-Ins*
  - In the *Add-Ins Available* box, select the check box next to *Analysis ToolPak*, and then click *OK*
  - If you see a message that tells you the *Analysis ToolPak* is not currently installed on your computer, click *Yes* to install it
  - When you load the *Analysis ToolPak*, the *Data Analysis* command is added to the *Tools* menu

Statistical Hypotheses

- Null hypothesis
  - H_0: $\mu_1 = \mu_2$
  - No difference between conditions
  - Any difference found in the data is due to chance

- Alternative hypothesis
  - H_1 or H_A: $\mu_1 \neq \mu_2$
  - There is a difference between conditions
  - The difference found is not due to chance

Statistical Significance

- *p*-value
  - Probability of obtaining your data **if H_0 is true**
  - The lower the *p*-value, the higher confidence you can have that the conditions are different

- If $p > .05$, the conditions are **not significantly different**

- If $p < .05$, the conditions are **significantly different**, and you can **reject H_0**

T-Test: Overview

- Compares the means of **two** sets of data to examine the effect of **one** independent variable on the dependent variable

- Two possible results
  - The data sets are significantly different
    - The IV had an effect on the DV
    - Differences observed probably due to the IV
  - The data sets are *not* significantly different
    - The IV had no effect on the DV
    - Differences observed probably due to chance

T-Test: One- vs. Two-Tailed

- One-tailed
  - The hypothesis is that Group A will have either a higher or lower mean than Group B
  - H_1: $\mu_1 > \mu_2$ **OR** H_1: $\mu_1 < \mu_2$
- Two-tailed
  - The hypothesis is only that Group A will have a different mean than Group B
    - The direction of the difference is not specified
  - H_1: $\mu_1 \neq \mu_2$

T-Test: Between-Groups

- Compares the means of **two different groups**
- Usually involves an experimental group and a control group
- *E.g.*, reaction times of males vs. females *or* group who drank beer vs. group who didn't

T-Test: Between-Groups

- Enter each data set into a column in Excel

- Under *Data Analysis*, select ***t-Test: Two-Sample Assuming Equal Variances***
  - Type in the cell range for Group 1 in *Variable 1* and for Group 2 in *Variable 2*
  - Type in `0` (zero) for *Hypothesized Mean Difference* (the null hypothesis of no difference)
  - Type in `0.05` for *Alpha* (the *p*-value or significance cut-off)
  - Click *OK*

T-Test: Reporting Results

t-Test: Two-Sample Assuming Equal Variances

	Variable 1	*Variable 2*
Mean	169.625	150.25
Variance *(Standard deviation = square root of Variance)*	65.125	121.9286
Observations	8	8
Pooled Variance	93.52679	
Hypothesized Mean Difference	0	
df	14	
t Stat	4.006856	
P(T<=t) one-tail	0.000649	
t Critical one-tail	1.76131	
P(T<=t) two-tail	0.001298	
t Critical two-tail	2.144787	

T-Test: Reporting Results

- t(df) = xx.xx, p = .xx, two-tailed [or one-tailed]
- *E.g.*, A standard between-groups t-test at the α = .05 level was conducted to test for a difference in the average confidence scores between the two age groups. The young group (M = 169.63, SD = 8.07) had significantly higher scores than the old group (M = 150.25, SD = 11.04) [$t(14) = 4.01$, p = .001, two-tailed], suggesting…

T-Test: Repeated-Measures

- Compares the means of two different measures from **one group**
- Usually involves before-treatment and after-treatment measures
- *E.g.*, reaction times before drinking beer vs. after drinking beer

T-Test: Repeated-Measures

- Enter each data set into a column in Excel
  - Need the same number of data points per condition
- Under *Data Analysis*, select ***t-Test: Paired Two-Sample for Means***
  - Type in the cell range for Measure 1 in *Variable 1* and for Measure 2 in *Variable 2*
  - Type in `0` for *Hypothesized Mean Difference*
  - Type in `0.05` for *Alpha*
  - Click *OK*

One-Way ANOVA: Overview

- Compares the means of **two or more** sets of data to examine the effect of **one IV**

- Two possible results:
  - **At least one** of the data sets are significantly different from at least one of the other data sets
  - The data sets are *not* significantly different from each other

- Test (p-value) always two-tailed

One-Way ANOVA (Between-Groups)

- Enter each data set into a column in Excel
- Under *Data Analysis*, select ***ANOVA: Single Factor***
  - Type in the cell range `first cell of first column:last cell of last column` in *Input Range*
  - Select *Columns* in the *Grouped By* option
  - Type in `0.05` for *Alpha*
  - Click *OK*

ANOVA: Reporting Results

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Column 1	6	203	33.83333	22.16667
Column 2	6	186	31	26
Column 3	6	255	42.5	19.5

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	430.7778	2	215.3889	9.549261	0.002114	3.68232
Within Groups	338.3333	15	22.55556			
Total	769.1111	17				

ANOVA: Reporting Results

- F(df between groups, df within groups) = xx.xx, p = .xx
- *E.g.*, A one-way between-groups ANOVA at the α = .05 level was conducted to test for the effect of incentives on homework effort. The number of correct answers differed significantly across the three groups [$F(2, 15) = 9.55$, $p = .002$]…

ONE-WAY ANOVA (REPEATED-MEASURES)

- Enter each data set into a column in Excel
- Under *Data Analysis*, select ***ANOVA: Two-Factor Without Replication***
  - Type in the cell range `first cell of first column:last cell of last column` in *Input Range*
  - Type in `0.05` for *Alpha*
  - Click *OK*
- F(df columns, df error) = xx.xx, p = .xx

ONE-WAY ANOVA (REPEATED-MEASURES)

Anova: Two-Factor Without Replication

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
Row 1	3	32.9	10.96667	38.80333
Row 2	3	17.5	5.833333	2.573333
Row 3	3	37	12.33333	40.33333
Row 4	3	35.3	11.76667	31.85333
Row 5	3	20.5	6.833333	31.96333
Column 1	5	51.3	10.26	34.703
Column 2	5	68.4	13.68	13.317
Column 3	5	23.5	4.7	0.285

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	107.584	4	26.896	2.512588	0.12452	3.837853
Columns	205.4173	2	102.7087	9.594906	0.007494	4.45897
Error	85.636	8	10.7045			
Total	398.6373	14				

ANOVA: Follow-Up Tests

- If ANOVA is significant, then need to compare each pair of groups or conditions to see where the difference is coming from

- Bonferroni correction
  - (Between-groups or repeated-measures) *t*-test on each pair of groups/conditions
  - For all *t*-tests: *p*-value = .05 / **no. of t-tests**

- Tukey post-hoc test
  - Most popular and acceptable follow-up test
  - But basic data analysis with Excel can't do it

ANOVA: Follow-Up Tests

- Reporting after main ANOVA results
  - …Multiple *t*-tests with a Bonferroni correction (α = .05/3 = .017) revealed that the pride group got significantly more correct answers than both the no incentive [$t(10) = -3.29$, p = .008, two-tailed] and toy groups [$t(10) = -4.18$, p = .002, two-tailed], suggesting…

Standard Error of the Mean

- Shows the uncertainty in how your sample mean represents the actual population mean
- The larger the sample size, the smaller it is
- The smaller it is, the more accurate the mean

- `=stdev(first cell of column:last cell of column)/sqrt(count(first cell:last cell))`
  - *E.g.*, `=stdev(b2:b17)/sqrt(count(b2:b17))`

- If two groups' standard error bars don't overlap, then they are significantly different

Graphs (Excel 2007)

- Insert → Charts → Column → 2-D Column
- Right-click on blank chart → Select Data
- Chart Data Range: Select all the group means
- Chart Tools → Layout → Analysis → Error Bars
  - More Error Bars
    - Display → Direction → Both
    - Error Amount → Custom → Specify Value
      - Positive Error Value: Select all standard errors (SEM)
      - Negative Error Value: Select all standard errors again
      - *Make sure to select in the same order as the means*
- Change labels, legends, color, etc. if necessary

GRAPHS: REPORTING RESULTS

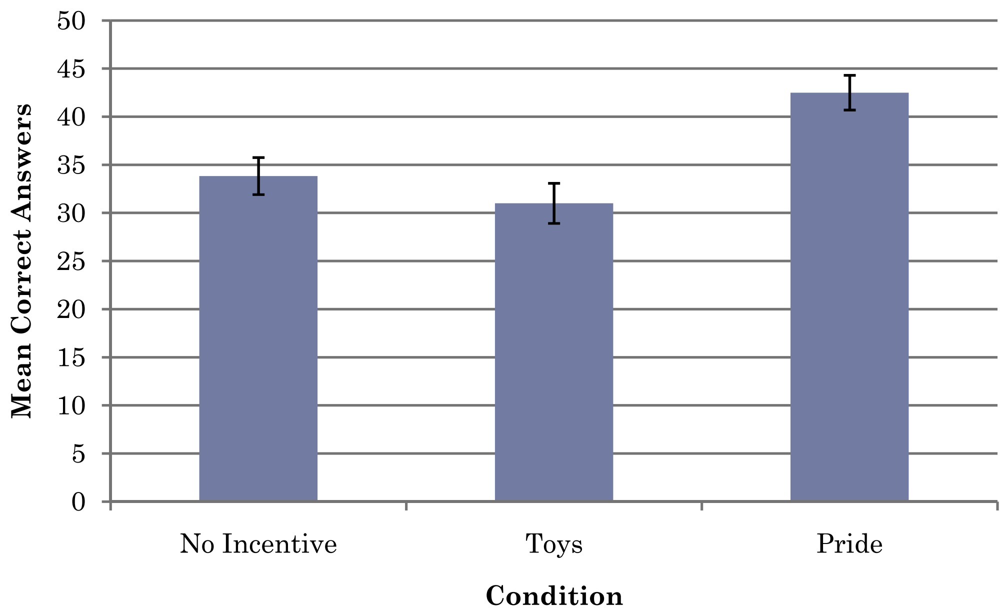


Figure 1. Mean number of correct answers for the different groups. Error bars represent the standard errors of the means.