

# STAT 235 - Lab 5



# Contact Info

- Email: [jmarley@ualberta.ca](mailto:jmarley@ualberta.ca)
- Office: CCIS L1-195
- Schedule: Tuesday 6pm, Wednesday 6pm and Friday 2pm
- Available labs to you: CAB 331, 335, 341, 345

# Things to Remember

- All assignments are typed
- Assignments must have a cover page WITHOUT your student ID (seriously, don't put it on there)
- Late assignments will not be accepted
- Caption/header on every figure and table.
- [www.stat.ualberta.ca/statslabs](http://www.stat.ualberta.ca/statslabs) for more info

## Just a note...

I (Jessa) am very open to discussing your grades from your assignments. Please if you have any concerns or questions, come talk to me (not your profs!) since I was the one who marked them. Sometimes I make counting mistakes (it happens) and I'm MORE than happy to change a grade if warranted.

# How to Caption/Header

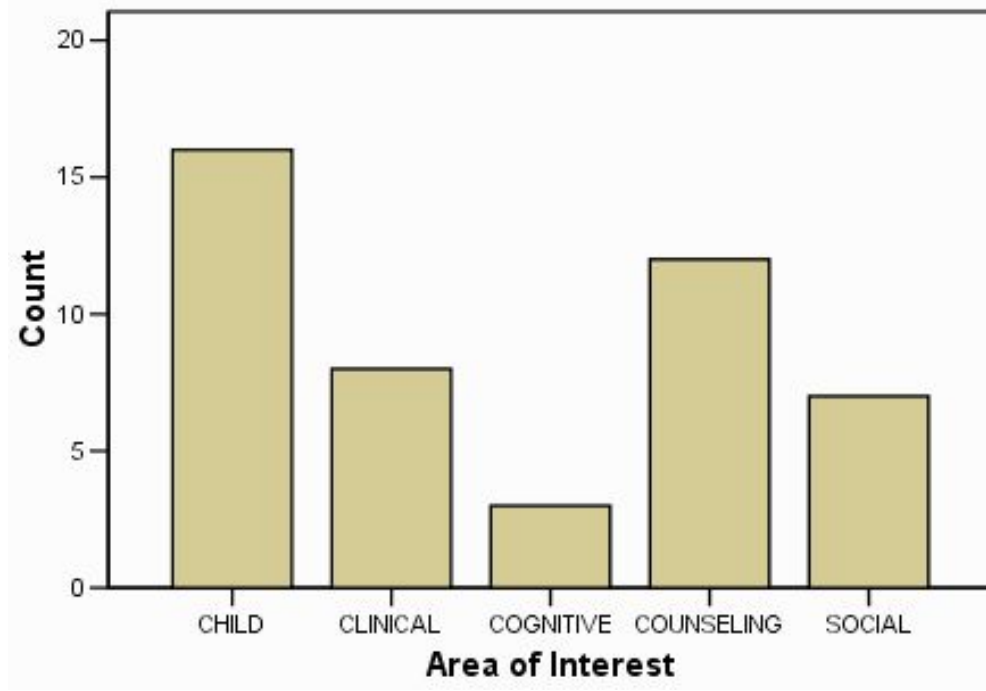


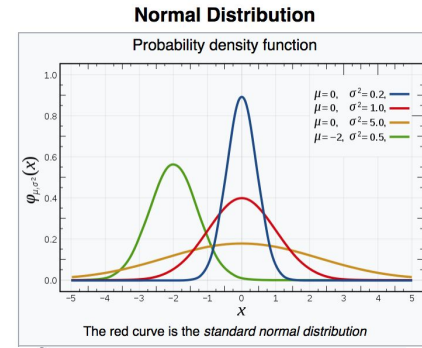
Figure 1: Histogram of career interested for students in 2008

# Lab 1 - Statistics

1. Histogram: Show the frequency (or count) for each bin.
2. Summary statistics: Mean (or average), standard deviation ( $\sigma$ ) and variance ( $\sigma^2$ ).
3. Quartiles: First = middle of smallest and median, Second = median, third = middle of median and highest. Splits data into four groups.
4. Mean change:  $\mu_2 - \mu_1$  change in means
5. Scatter plots: Visualization of your data, compares the data under two conditions (x and y). Typically the response goes on the y-axis, i.e. how does x influence y?
6. Interquartile Range: Spread between third and first quartile.  $Q3 - Q1$

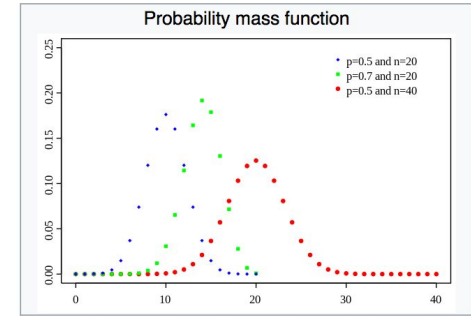
# Lab 2 - Statistics

1. Normal Distribution: Bell shaped, continuous, positive and negative, it has two parameters ( $\mu$  = mean,  $\sigma$  = standard deviation). Note that this distribution is symmetric.
2. Mean: For the normal distribution, the mean occurs at the maximum value.
3. Standard Deviation: For the normal distribution this describes the spread of the distribution.
4. Random Number Generation: Gives random numbers based on a desired distribution. Note that seeds results in the same random numbers. (Ask me about this if you're curious)
5. Binomial distribution: DISCRETE, looks at number of successes



$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

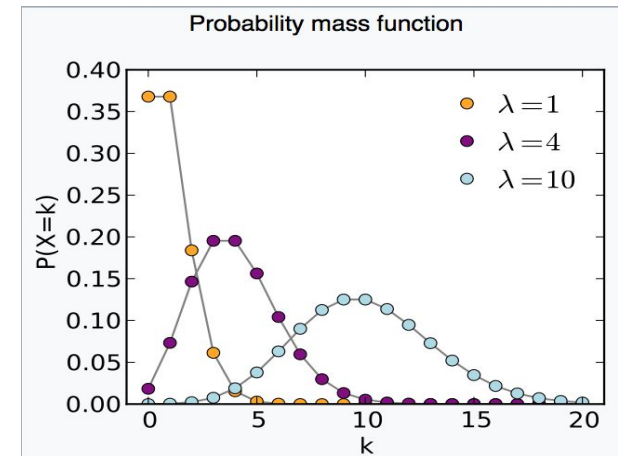
## Binomial distribution



$$\binom{n}{k} p^k (1-p)^{n-k}$$

# Lab 3 - Statistics

1. Poisson Distribution: Discrete, positive, one parameter ( $\lambda$ )
2. Lambda ( $\lambda$ ): Expected number of occurrences.
3. Central limit theorem: If a random sample of size  $n$  is drawn from a population with mean  $\mu$  and variance  $\sigma^2$ , the sample mean approximately follows a normal distribution with mean  $\mu$  and variance  $\sigma^2/n$  if the sample size is sufficiently large. Also averages of samples converge to normal distributions.



$$\frac{\lambda^k e^{-\lambda}}{k!}$$

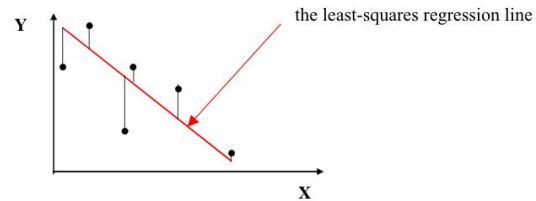


# Lab 4 - Statistics

1. Confidence intervals: the range of which we are #% certain. Note that a wider range means more uncertainty.
2. Margin of error: provides information about the accuracy of the estimate of the population parameter.
3. Two sided (tailed) test: checks if alternate hypothesis is not equal
4. One sided (tailed) test: checks if alternate hypothesis is greater or lesser. Need to divide two-tailed p-values by 2.
5. T-test: Comparison of two means, requires normality.
6. P-value: the probability that your alternate hypothesis is wrong, typical use  $\alpha=0.05$  or  $=0.01$

# Lab 5 - Statistics

1. Correlation: Determines if there is a relationship between variables (not whether the relationship is linear),  $r$  is the strength of the relationship,  $-1 \leq r \leq 1$ ,  $0$  = no relationship,  $1$  = positive relationship,  $-1$  = negative relationship
2. Linear Regression: linear relationship between independent variable and dependent variable.  $Y = \beta_0 + \beta_1 X_1 + \varepsilon$
3. Multiple regression:  $Y = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n + \varepsilon$
4. Least squares regression line: minimizes distance from points to line.



5. Residual plot: plot of distances from points to line, should not have a pattern.

# Lab 1 - Excel

1. Histogram: Add ins (under Insert) -> Manage other Add-ins -> Go... -> Check "Analysis Toolpak" -> Data Analysis (under Data) -> Histogram -> Choose data and check "Chart Output"
2. **Summary Statistics: (If "Analysis Tool Pack" is already added in) Data Analysis (under Data) -> Descriptive Statistics -> Choose data and check "Summary Statistics"**
3. Quartiles: Insert Formula (under Formulas) -> formula is "QUARTILE.INC" -> Choose data and quartile number.
4. Mean Change: Can be found using a formula, one cell minus the one above. You won't have a value for the first one.
5. Scatter Plot: Scatter (under Insert) -> Right click and "Select Data" to change the plotted data
6. Interquartile Range:  $IQR = Q3(\text{third quartile}) - Q1(\text{First quartile})$

## Lab 2 - Excel

1. *Normal Distribution: Formulas -> Insert Function -> there are several function for the normal dist. (NOT NECESSARY FOR THIS LAB)*
2. Mean: Can be found using summary statistics.
3. Standard Deviation: Can be found using summary statistics.
4. Random Number Generation: Data -> Data Analysis -> Random Number Generation -> number of variables = #, Number of random variables = #, Distribution = "dist", seed = #, mean = #, st. deviation = #
5. *Binomial Distribution: Formulas -> Insert Function -> There are several functions available. (NOT NECESSARY FOR THIS LAB)*

# Lab 3 - Excel

1. Poisson Distribution: Don't worry about it, it's given to you.
2. Lambda ( $\lambda$ ): This is just entered by you this time, not fitting a distribution.

# Lab 4 - Excel

1. Confidence intervals: Insert function -> Confidence Norm -> alpha = #, standard deviation = #, size = #)
2. Margin of error: Given to you in lab, also in summary statistics.
3. T-test: Data Analysis -> t Test (three versions, depends on what you need)

# Lab 5 - Excel

1. Correlation: Data Analysis -> Correlation -> put in correct data
2. Multiple Regression: Data Analysis -> Regression -> Only choose "Constant is zero" if you want line to go through (0,0).
3. Residual Plot: Select "Residual Plots" in the regression window

# Lab 5 - Data

The data for this lab was collected to determine how jet engine thrust is impacted by various factors.

THRUST - Thrust of a jet turbine engine

RATE - Fuel flow rate

EXTEMP - Exhaust temperature

AMBTEMP - Ambient temperature at time of test



# Lab 5 - Guidelines

Q1. a. **OUTPUT: 3 Scatterplots, thrust vs. Fuel flow rate, thrust vs. exhaust temperature, thrust vs. Ambient temperature.** What should be on the y-axis?

Q1. b. Look at the center of the “clouds”. Are they linear? Strength of linear relationship?

Q2. a. **OUTPUT: Correlation matrix.**

Q2. b. Highest correlation, lowest (note absolute value)? Does this match the conclusions from Q1?

Q3. Define the model as an equation, use the basic equation of regression. Assumptions of regression?

**Q4. OUTPUT: Regression output from excel, not necessary/being marked.**

Q4. a. Equation of regression with coefficient values. Determine influential points from scatter plot in Q1. Determine outlier residuals from residual plot.

Q4. b. Standard deviation? Compare to average.

Q4. c. Percent of variation? What might explain variation?

Q4. d. Hypotheses? What kind of test can you use? Values of test statistics? P-value? Null distribution? Conclusion?

Q4. e. Interpret coefficients. Confidence interval?

Q4. f. Predicted value? Residual?

Q4. g. **OUTPUT: Residual plot.** Interpret the plot, are the assumptions met?

Q5. **OUTPUT: Regression output.** Optional.

Q5. a. Equation with numerical values of coefficients. Determine influential points from scatter plot in Q1. Determine outlier residual from residual plot.

Q5. b. Standard deviation? Compare to average.

Q5. c. Percent of variation?

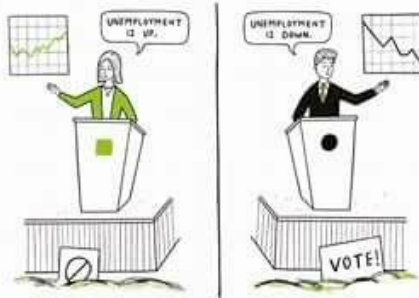
Q5. d. Hypotheses? Distribution? Test statistic? P-value? Conclusion?

Q5. g. **OUTPUT: Residual Plot.** Interpret the plot. Are the assumptions met?

Q6. What's the best predictor? S = model standard deviation

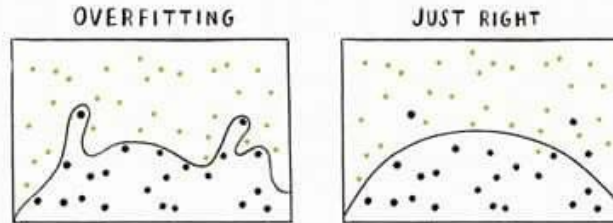
Thrust vs.	$R^2$	s	t-statistic	p-value
Rate	###	###	###	###
Extemp	###	###	###	###

# Data Fallacies to Avoid (aka how to be a bad statistician)



## CHERRY PICKING

Selecting results that fit your claim and excluding those that don't.



## OVERFITTING

Creating a model that's overly tailored to the data you have and not representative of the general trend.



## DATA DREDGING

Repeatedly testing new hypotheses against the same set of data, failing to acknowledge that most correlations will be the result of chance.

# Data Fallacies to Avoid (aka how to be a bad statistician)



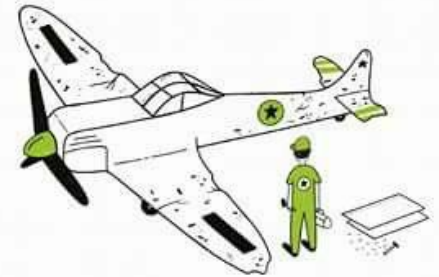
## COBRA EFFECT

Setting an incentive that accidentally produces the opposite result to the one intended. Also known as a Perverse Incentive.



## PUBLICATION BIAS

Interesting research findings are more likely to be published, distorting our impression of reality.



## SURVIVORSHIP BIAS

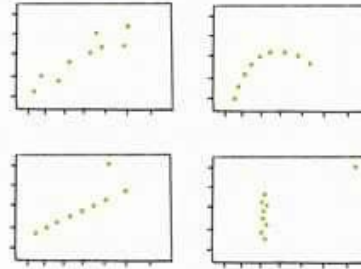
Drawing conclusions from an incomplete set of data, because that data has 'survived' some selection criteria.

# Data Fallacies to Avoid (aka how to be a bad statistician)



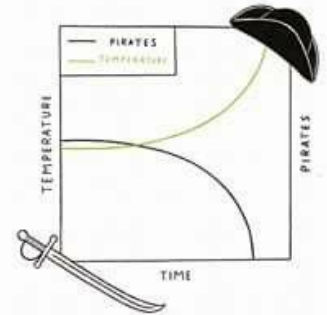
## SAMPLING BIAS

Drawing conclusions from a set of data that isn't representative of the population you're trying to understand.



## DANGER OF SUMMARY METRICS

Only looking at summary metrics and missing big differences in the raw data.



## FALSE CAUSALITY

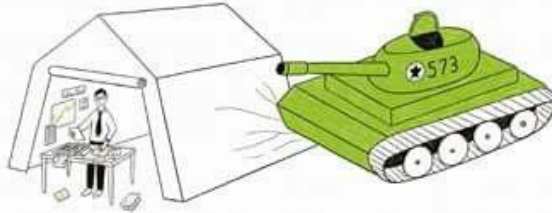
Falsely assuming when two events appear related that one must have caused the other.

# Data Fallacies to Avoid (aka how to be a bad statistician)



## REGRESSION TOWARDS THE MEAN

When something happens that's unusually good or bad, it will revert back towards the average over time.



## MCNAMARA FALLACY


Relying solely on metrics in complex situations and losing sight of the bigger picture.



## GAMBLER'S FALLACY

Mistakenly believing that because something has happened more frequently than usual, it's now less likely to happen in future (and vice versa).

# Data Fallacies to Avoid (aka how to be a bad statistician)



APPLICATION SUCCESS RATE

	MALE	FEMALE
SUBJECT 1	19 % (148 of 1200)	15 % (170 of 1800)
SUBJECT 2	50 % (400 of 800)	51 % (102 of 200)
TOTAL	28 % (548 of 2000)	19 % (372 of 2000) ??

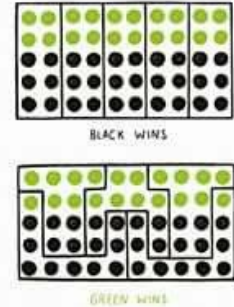
## SIMPSON'S PARADOX

When a trend appears in different subsets of data but disappears or reverses when the groups are combined.



## HAWTHORNE EFFECT

The act of monitoring someone can affect their behaviour, leading to spurious findings. Also known as the Observer Effect.



## GERRYMANDERING

Manipulating the geographical boundaries used to group data in order to change the result.