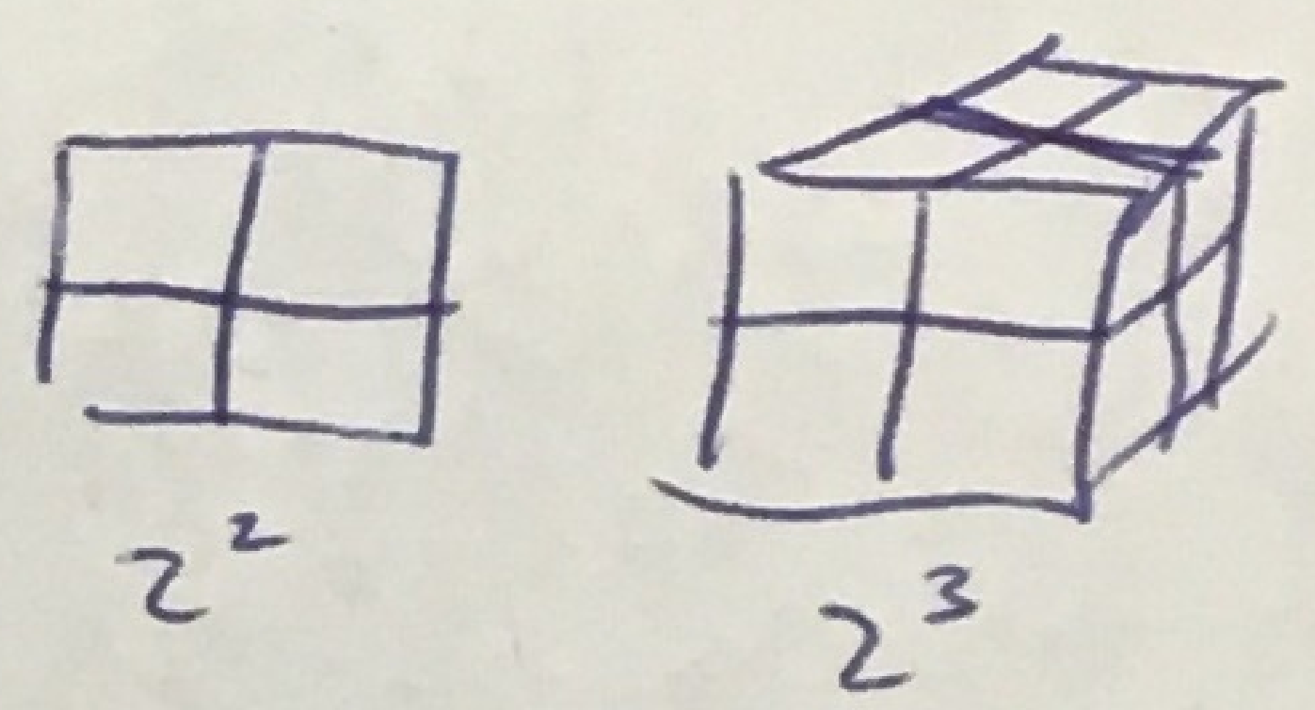
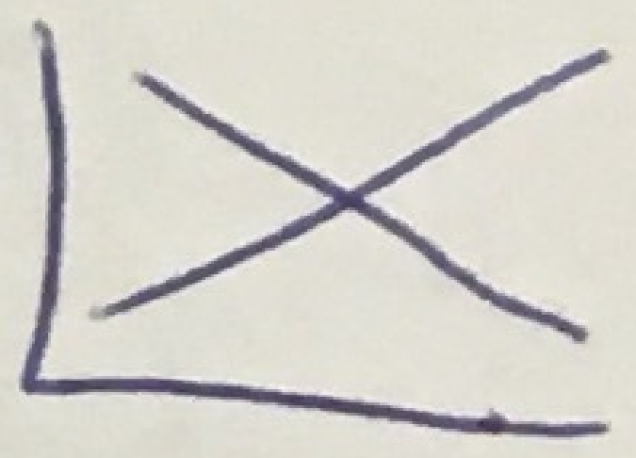
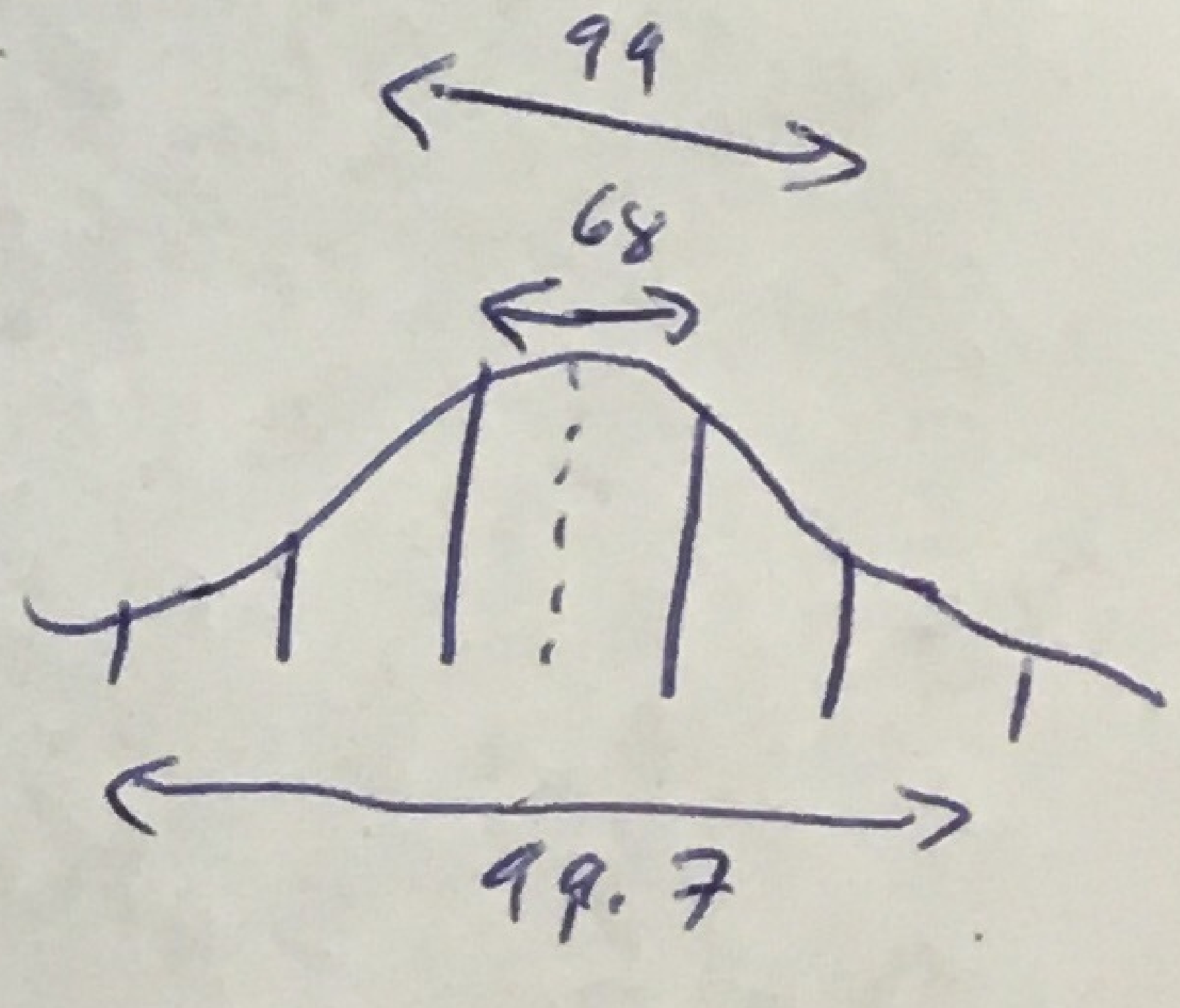
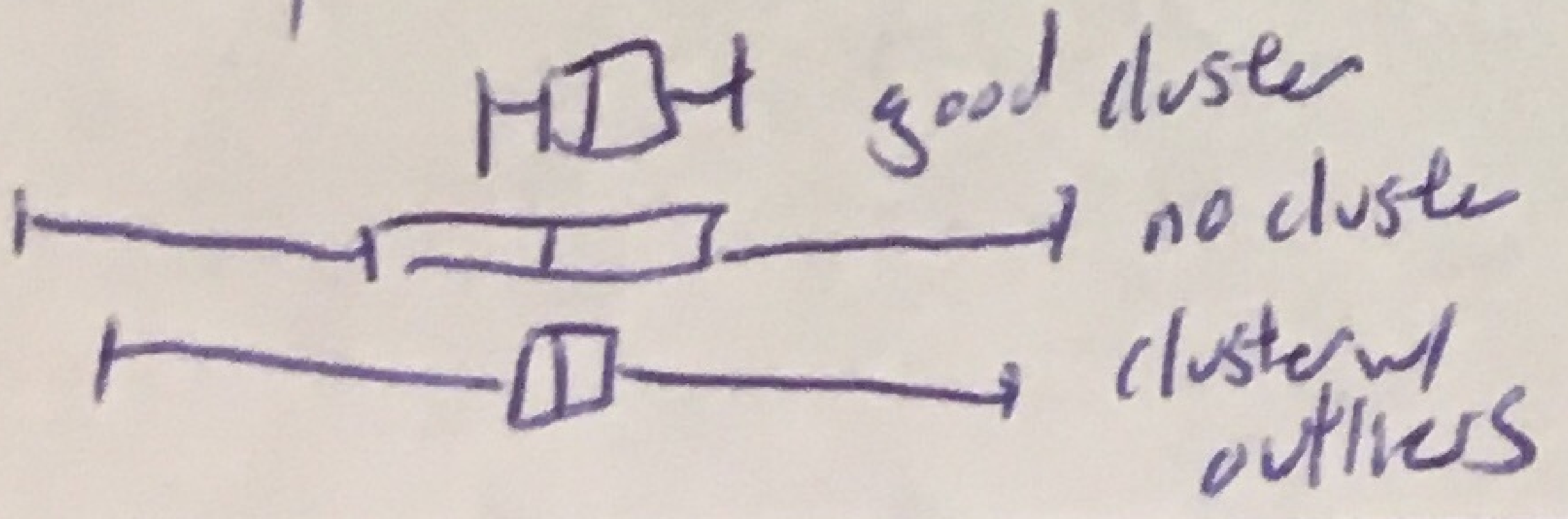
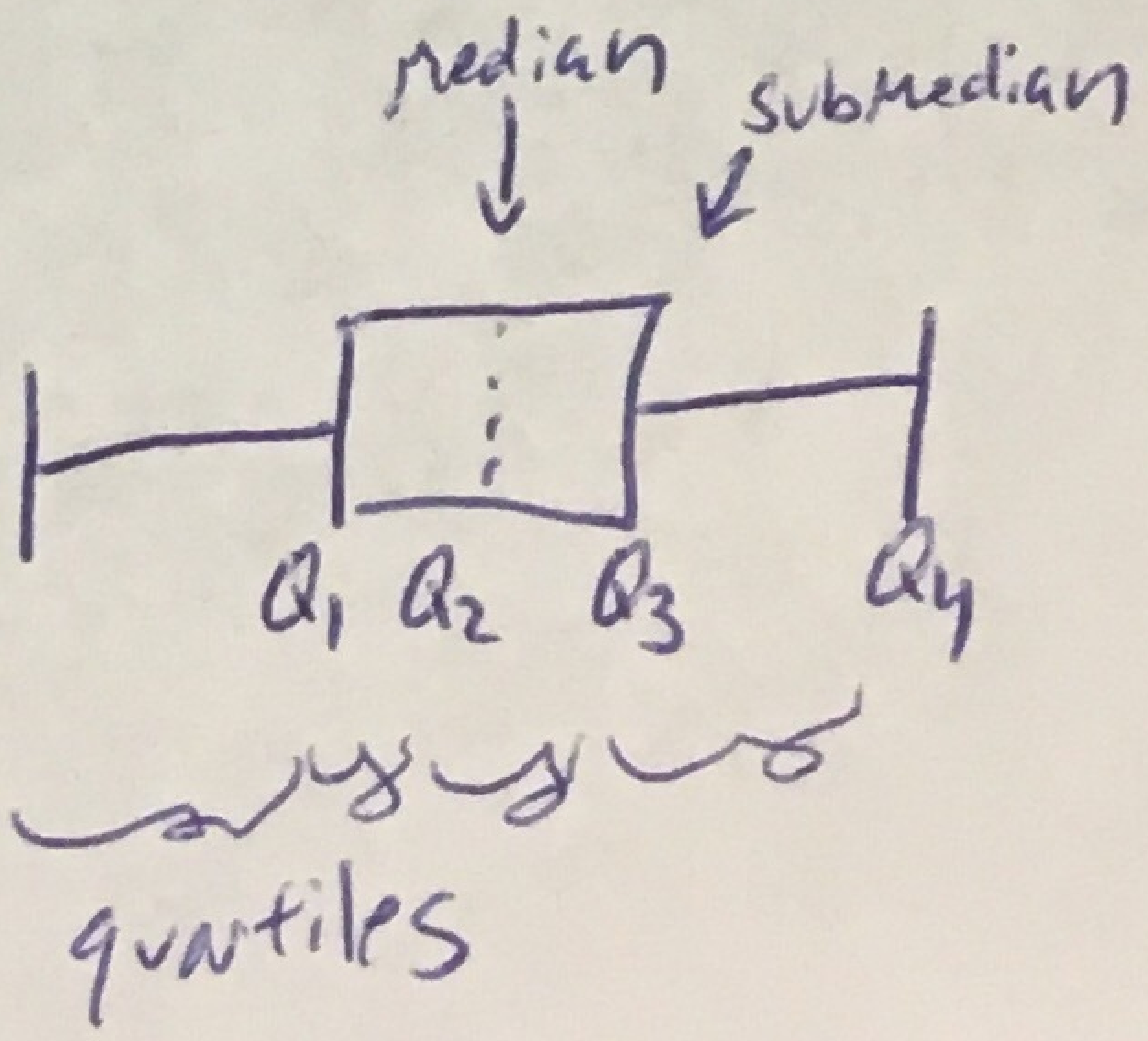
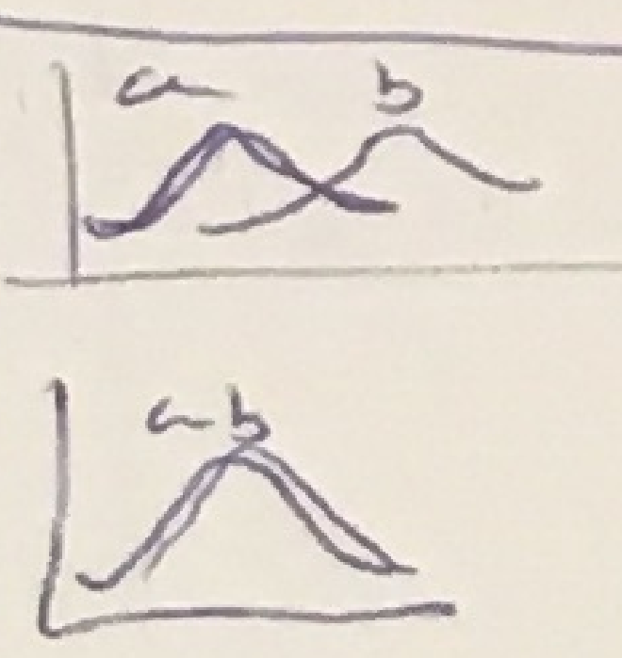
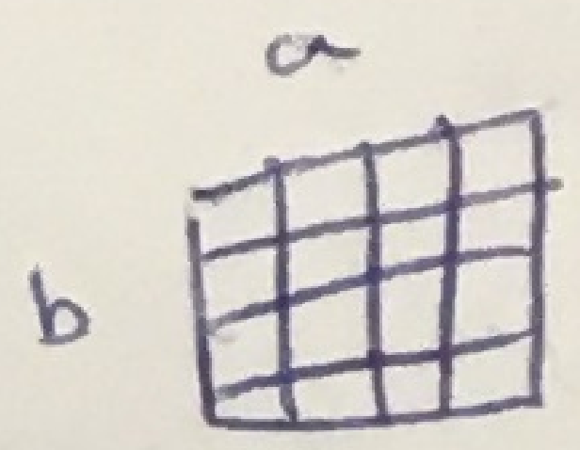
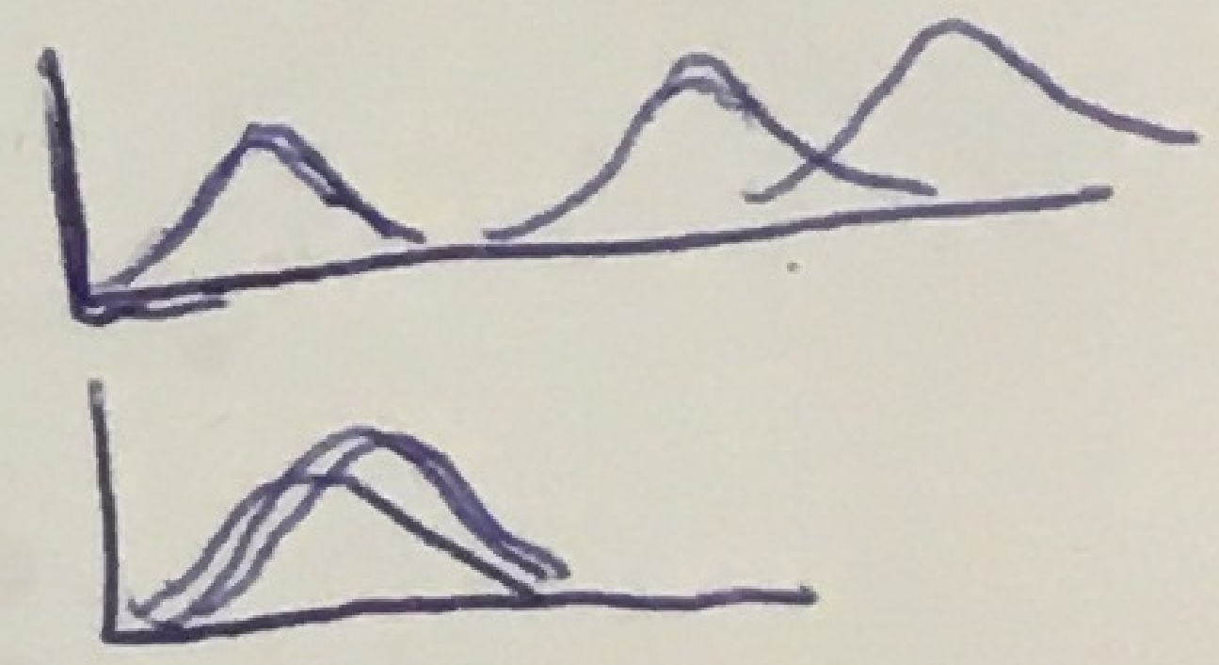
Statistics: Basic Definitions

Thing	Description	Formula
standard deviation	$\sqrt{\text{variance}}$	$\sqrt{\frac{\Sigma(\text{point}-\text{mean})^2}{n-1}}$
Variance	analysis of spread from mean	$\frac{\Sigma(\text{point}-\text{mean})^2}{n-1}$
Confidence interval	the range of values that let you be some % certain that the real value falls in there for some measure	depends on significance level, std dev, sample size
Factor Design	Investigate effect of 2 or more independent variables on one dependent variable	 each IV is a "factor" each factor has "levels"
Interaction Effects	Is response additive or not? Test with ANOVA and regression	use visualization: look for nonparallel lines 
(Main effect)	effect of IV on DV averaged across other IVs	
Tests for Normality	K-squared back of envelope 68-95-99.7 rule	
Calculating Outliers	IQR: width of box (Q_3-Q_1) outlier is any point more than 1.5 x IQR from Q_1 or Q_3	
Box + Whisker	based on medians whiskers show spread 	 median submedian Q_1 Q_2 Q_3 Q_4 quartiles

Statistics : Significance Tests

Test	Assumptions	Use	
t-test	normal distribution Variance of populations are same (student) or ~~diffrent~~ ~~unknown~~ diffrent (welch)	tell if 2 sets are significantly different	
Z-test	normal distribution population Variances are known large sample size (>30) know std dev	tell if 2 sets are significantly different	if small sample, or unknown variance use a t-test
Chi-squared	Chi-squared distribution Variables are categorical (nonparametric)	tell if there is an association between ~~multiple~~ two variables (e.g. job type and residence area	
ANOVA	independence normal distributions Variances same across groups	generalize t-test to more than 2 groups; test 3 or more means for significance	limits false positives 
Mann-Whitney U test	nonparametric (does not have to be normal) independent samples	2 populations, test for significant difference	almost as good as t-test on normal distros For dependent Samples, use Wilcoxon Signed rank
Spearman's Rank Correlation Coefficient	nonparametric both continuous and discrete variables	rank correlation: statistical dependence between the ranking of two variables	Pearson assumes linear; Spearman just measures monotonicity

F-test

F-distribution

Compare fits to find best Model; usually after least-squares fit

ANOVA

$$F = \frac{\text{explaind variance}}{\text{unexplained variance}}$$

this is used inside an ANOVA, or that a regression is a good fit