Lecture 4
Segment 3

Interpretation of correlations

Correlations 3

- Important topics
  - – Validity of a correlation-based argument
  - – Reliability of a correlation

Validity

- Assumptions underlying correlation analyses:
  - Normal distributions for X and Y
  - Linear relationship between X and Y
  - Homoskedasticity

Validity

- The validity of any argument made on the basis of a correlation analysis depends on these assumptions

Validity

- Assumptions underlying correlation analyses:
  - Normal distributions for X and Y
    - How to detect violations?
      - Plot histograms and run descriptive statistics

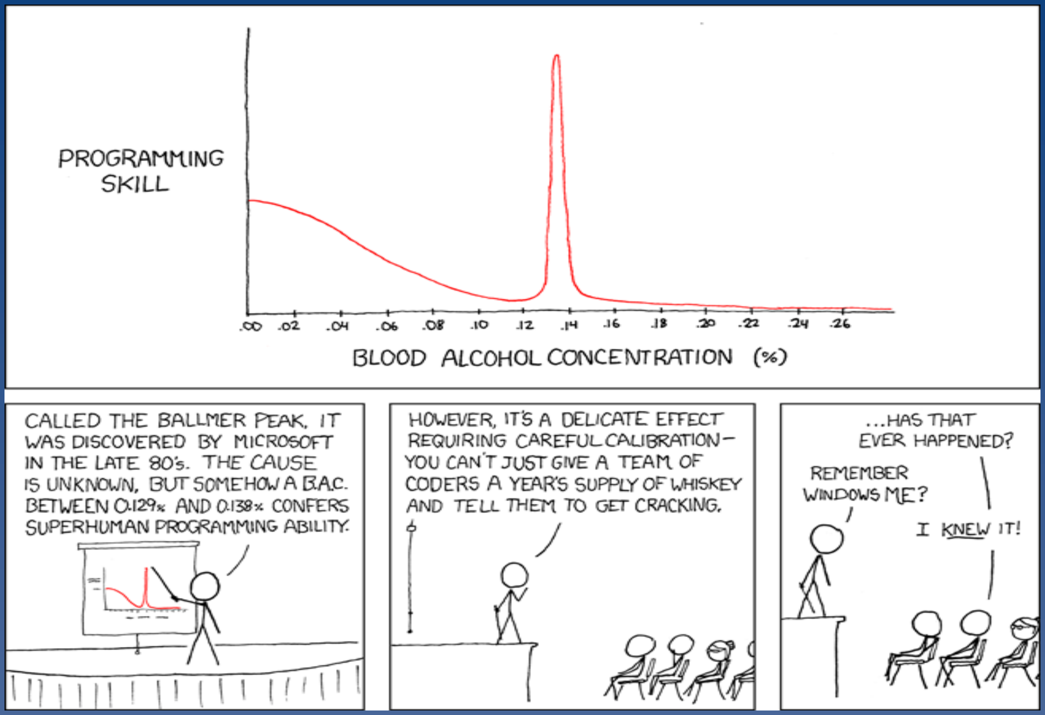
Validity

- Assumptions underlying correlation analyses:
  - Linear relationship between X and Y
    - How to detect violations?
      - Examine scatterplots (see following examples)
      - Plot a histogram of residuals (more on this later)

Validity

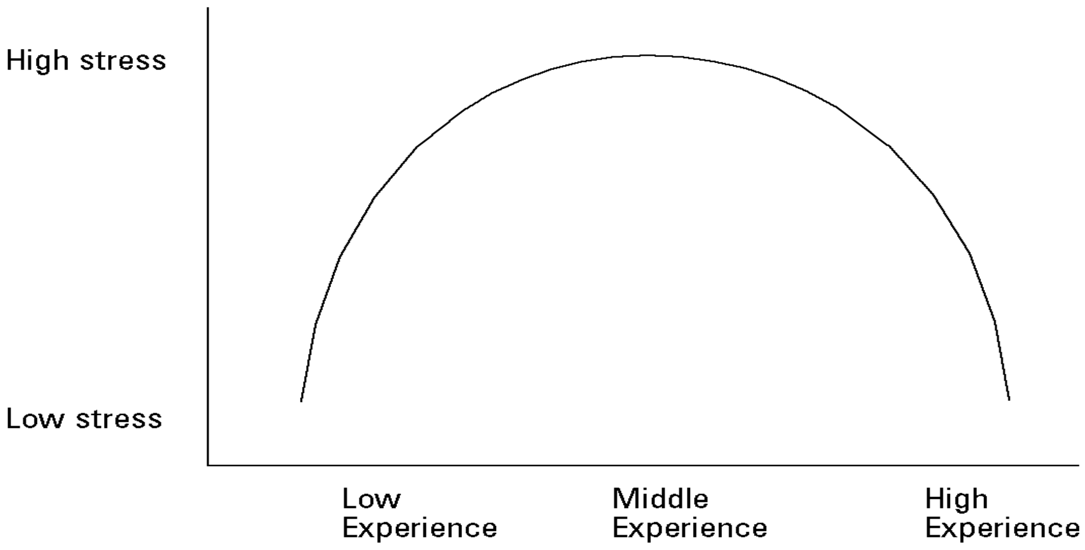
- Assumptions underlying correlation analyses:
  - Homoskedasticity
    - How to detect violations?
      - Examine scatterplots (see following examples)
      - Plot a histogram of residuals (more on this later)

Non-linear relation: Fun example


PROGRAMMING SKILL
.00 .02 .04 .06 .08 .10 .12 .14 .16 .18 .20 .22 .24 .26
BLOOD ALCOHOL CONCENTRATION (%)
CALLED THE BALLMER PEAK, IT WAS DISCOVERED BY MICROSOFT IN THE LATE 80's. THE CAUSE IS UNKNOWN, BUT SOMEHOW A B.A.C. BETWEEN 0.129% AND 0.138% CONFERS SUPERHUMAN PROGRAMMING ABILITY.
HOWEVER, IT'S A DELICATE EFFECT REQUIRING CAREFUL CALIBRATION – YOU CAN'T JUST GIVE A TEAM OF CODERS A YEAR'S SUPPLY OF WHISKEY AND TELL THEM TO GET CRACKING.
...HAS THAT EVER HAPPENED?
REMEMBER WINDOWS ME?
I KNEW IT!

Non-linear relation: Serious example

Figure 2 Schematic representation of quadratic relationship between stress and experience extrapolated from data in Table I



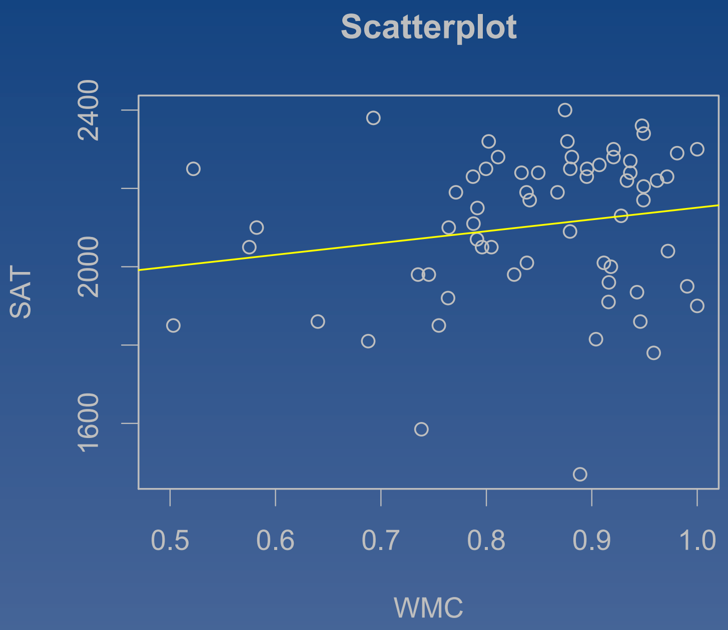
Homoskedasticity?!

- In a scatterplot the distance between a dot and the regression line reflects the amount of prediction error
    - In the baseball example, look at one baseball player, using the regression line to guide you, does his score on X accurately predict his score on Y?
        - If so then the distance will be small
        - If not then the distance will be large

Homoskedasticity?!

- Homoskedasticity means that the distances (the errors, or residuals) are not related to the variable plotted on the X axis (they are not a function of X)
- This is best illustrated with a scatterplot

R scatterplot: plot(SAT~WMC)



$r = .22$

Validity

- Validity of correlation-based arguments depends on several assumptions
  - Normal distribution in X and Y
  - Linear relationship between X and Y
  - Homoskedasticity

Reliability

- Reliability of a correlation
  - Does the correlation reflect more than just chance covariance?
  - One approach to this question is to use NHST

NHST

- Null Hypothesis Significance Testing (NHST)
  - H_0 = null hypothesis: e.g., r = 0
  - H_A = alternative hypothesis: e.g., r > 0

NHST

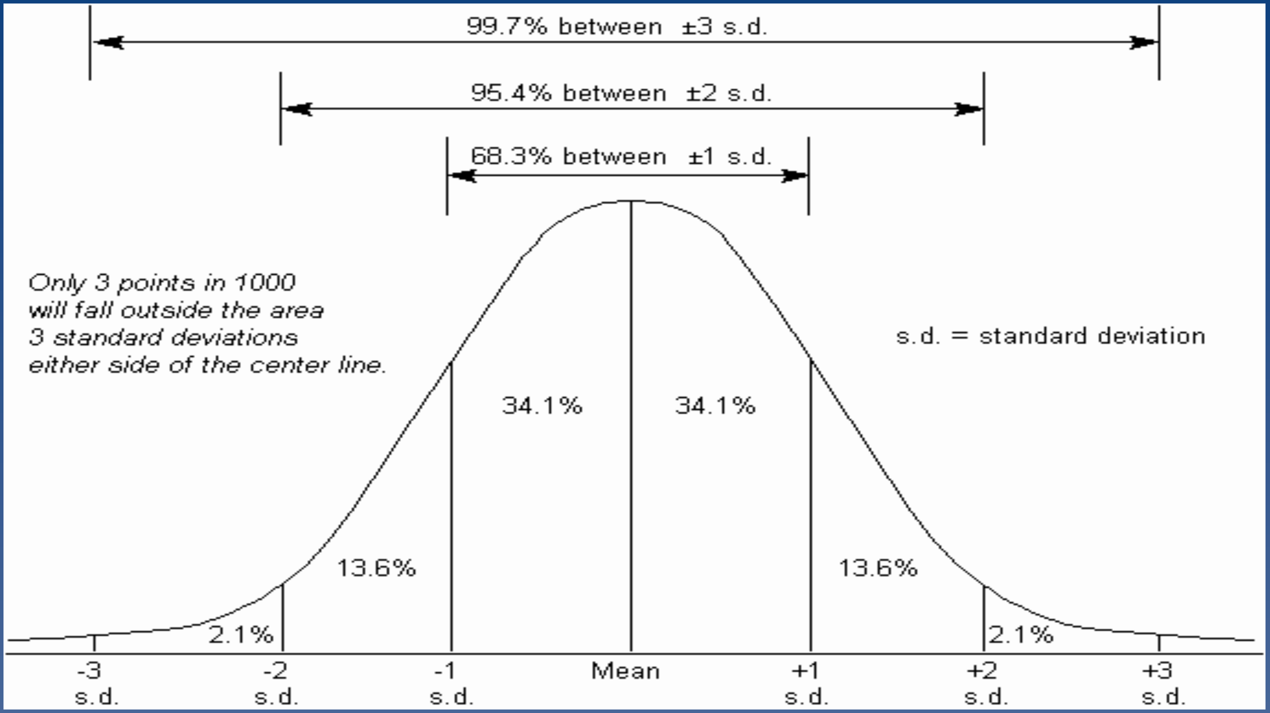
- Null Hypothesis Significance Testing (NHST)
  - Assume H_0 is true, then calculate the probability of observing data with these characteristics, given that H_0 is true
    - Thus, $p = P(D|H_0)$
    - If $p < \alpha$ then Reject H_0, else Retain H_0

NHST

Experimenter Decision

		Retain H_0	Reject H_0
Truth	H_0 true	Correct Decision	Type I error (False alarm)
	H_0 false	Type II error (Miss)	Correct Decision

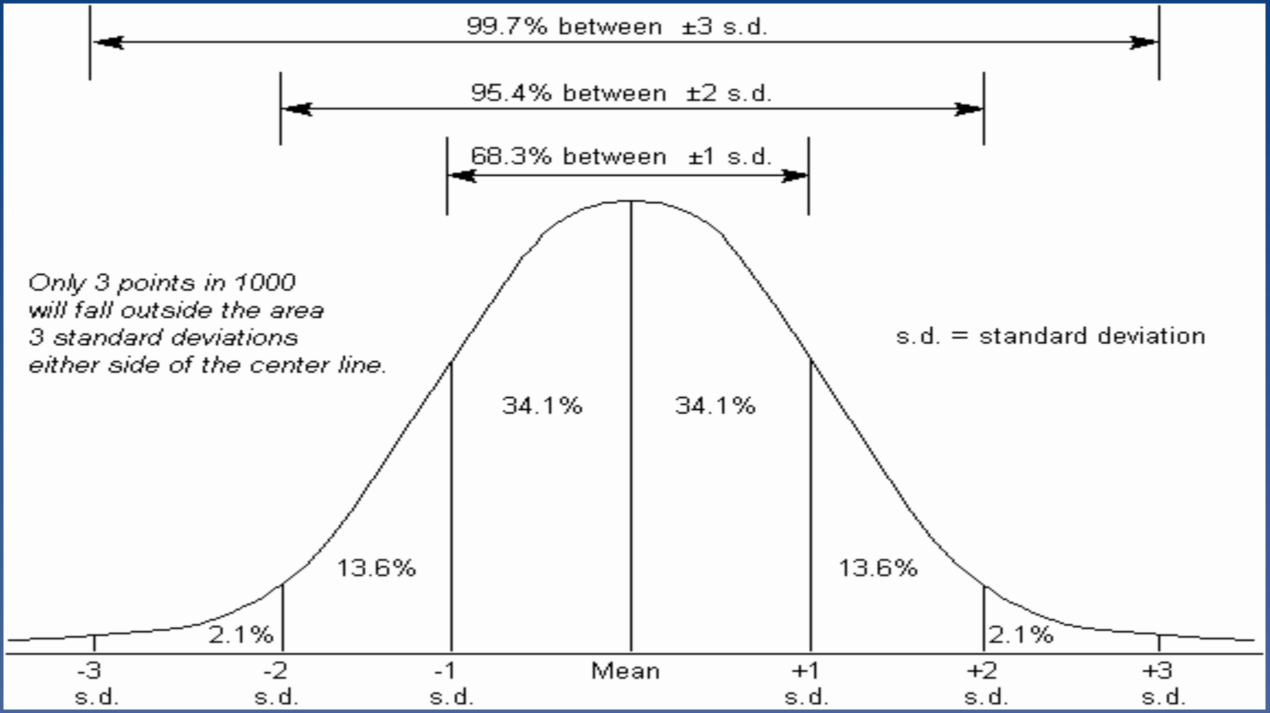
The normal distribution



NHST

		Experimenter Decision	
		Retain H_0	Reject H_0
Truth	H_0 true	$p = (1 - \alpha)$	$p = \alpha$
	H_0 false	$p = \beta$ (1 - POWER)	$p = (1 - \beta)$ POWER

The normal distribution



NHST

- $p = P(D|H_0)$
- Given that the null hypothesis is true, the probability of these, or more extreme data, is p
  - NOT: The probability of the null hypothesis being true is p
  - In other words, $P(D|H_0) <> P(H_0|D)$

NHST can be applied to:

- *r*
  - Is the correlation significantly different from zero?
- *r*1 vs. *r*2
  - Is one correlation significantly larger than another?

Correlations: Final note

- There are other correlation coefficients
  - Point biserial *r*
    - When 1 variable is continuous and 1 is dichotomous
  - Phi coefficient
    - When both variables are dichotomous
  - Spearman rank correlation
    - When both variables are ordinal (ranked data)

Correlations 3

- Important topics
  - Validity of a correlation-based argument
  - Reliability of a correlation

Image in slide 8 was retrieved from http://xkcd.com/323/

Image in slide 9 is from Moran, C. C. (1998). Stress and emergency work experience: a non-linear relationship. *Disaster Prevention and Management*, 7(1), 38 - 46

Image in slides 18 and 20 was retrieved from http://www.syque.com/quality_tools/toolbook/Variation/Image375.gif