Psychology 252: Statistical Methods for the Social Sciences (9/22/14)

Instructors: Benoît, Ewart, Rebecca, Caitie, Kara

Topics include: **GLM** (ANOVA, Regression) and **GLMM**, or Mixed Models

Texts: *Howell, Intros to R (the stat package)*

Handouts (4): Syllabus, HW-1; HO-1 on Coursework; plus lecture slides

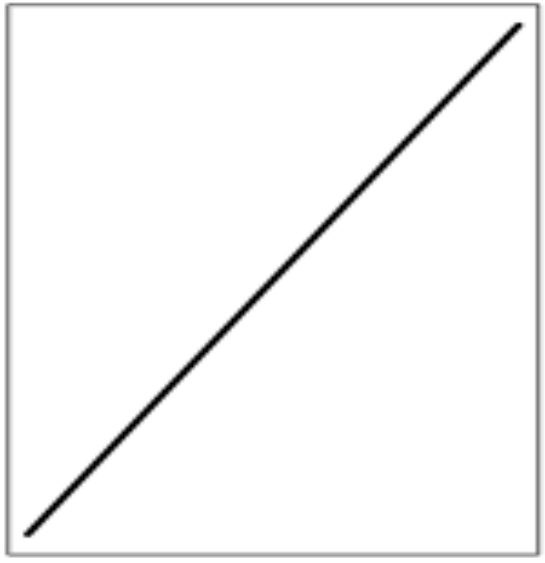
Work: Group work encouraged on **HW**, but **write up your own solutions**. **Quizzes** (2 in-class, 2 take-home) **must be own work**.

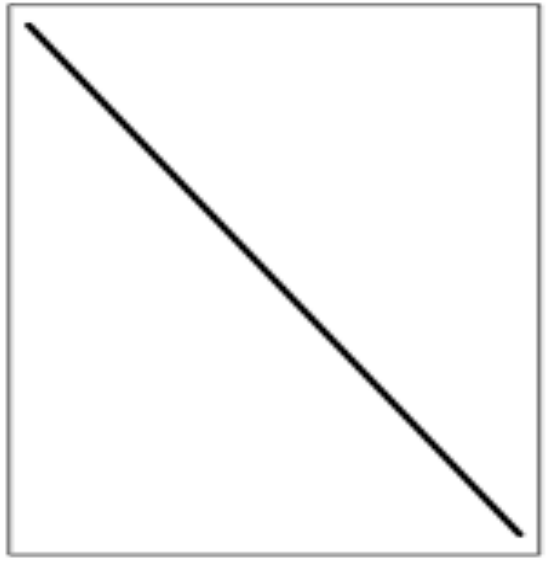
Secs, WTh; HW-1 and R

- Our goal is to understand Statistics; access to R packages facilitates this. Expertise in Stats and R is distributed, so we'll need to help each other. We hope last Friday's R Tutorial was helpful.
- For next 2 weeks or so, attention to R vs. Stats will still be greater in WTh Sections than in Lectures, based on ***'stutorial1/2/3.Rmd'*** in the Week 0 folder.
- HW-1, due 10/01, contains stats review & R material, and it is 'long'. Please start it soon.
- Relevant *.r scripts in Coursework, useful for Handouts & HWs. Learning R by imitation.

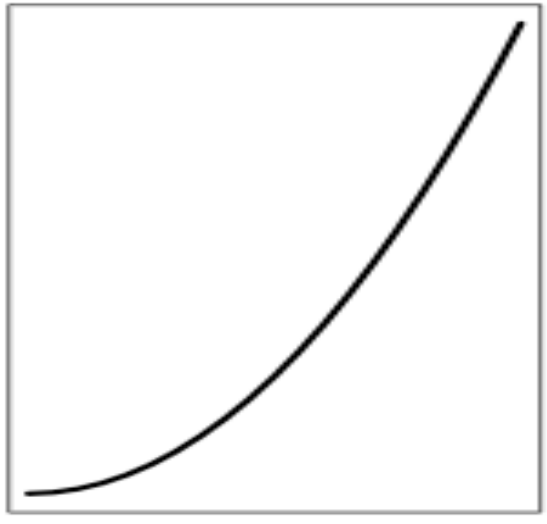
Packages in R

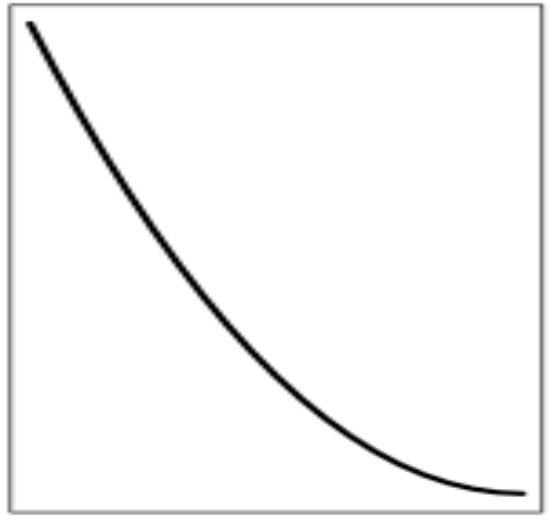
- R is a *freely* available language and environment for statistical computing and graphics; related to the commercial S-Plus; pun on '*S*', or eponym for original authors, *R*obert Gentleman & *R*oss Ihaka (U of Auckland).
- Widespread use around Stanford & the world – along with Matlab (Octave), Stata (Econ).
- R is powerful, flexible, provides access to 5870 (in last 5 years: 4769, 4048, 3250, 2534 and 1750!) special-purpose packages, e.g., car, psy, lme4, concord, date, ggplot2. **Time trend?**



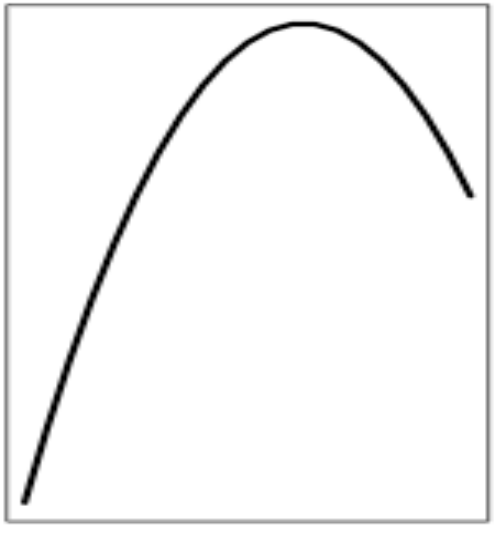


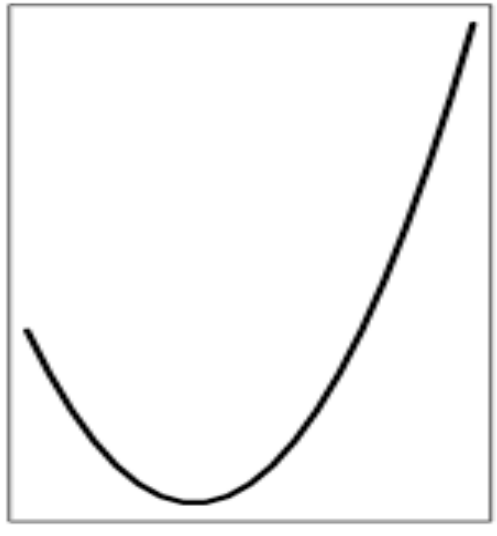
Possible time-trends





Possible time-trends





Growth in number of packages:

[Use R console to demo]

- Plot data, *n* versus *t:*

  t1 = 1:6; n1 = c(1750, 2534, 3250, 4048, 4769, 5870)

- Do linear regression, save `lm()` object for
  - graphing of regression line, and
  - estimating slope and intercept.
- Test for non-linearity with `poly(t, 2)`

```
rs1 = lm(n1 ~ t1)
plot(rs1); summary(rs1)
plot(t1, n1); abline(rs1)
rs2 = lm(n1 ~ poly(t1, 2))
summary(rs2)
```

Lifespan maturation and degeneration of human brain white matter (9/17/2014)

Jason D. Yeatman**, Brian A. Wandell & Aviv A. Mezer

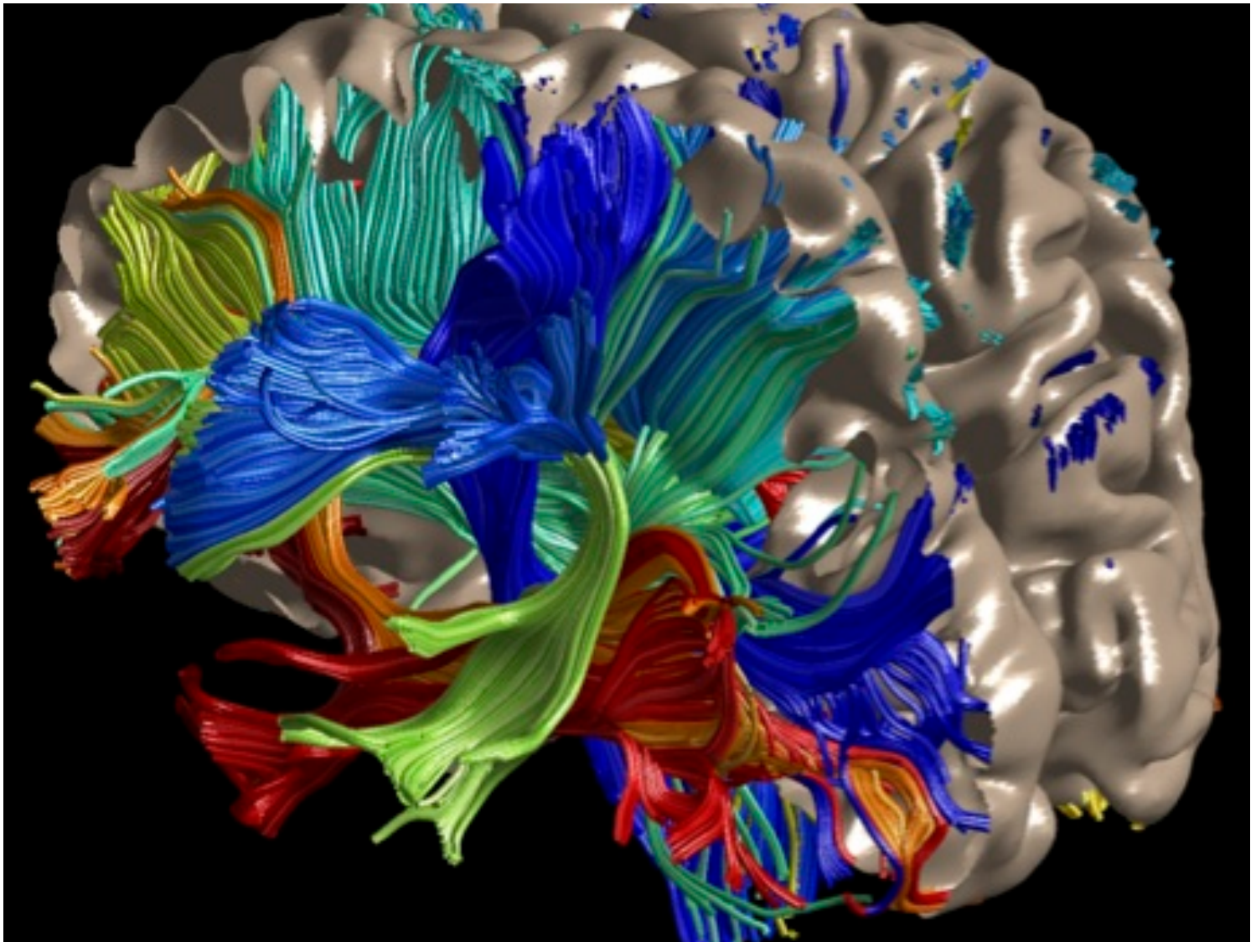
**** Former Psych 252 TA!**

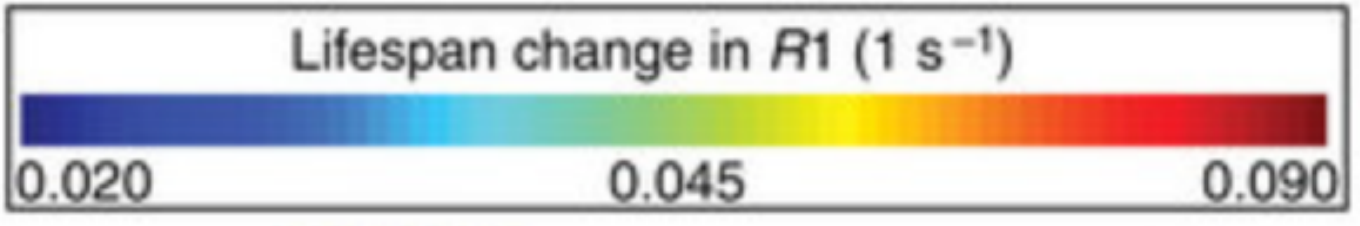
- White matter is composed of bundles of myelinated nerve cell processes (or axons), which connect various grey matter areas (the locations of nerve cell bodies) of the brain to each other, and carry nerve impulses between neurons. White matter plays a critical role in nearly every aspect of cognitive development, healthy cognitive function and cognitive decline in aging. Moreover, many psychiatric disorders—from autism to schizophrenia—are associated with white-matter abnormalities.
- Myelin acts as an insulator, increasing the speed of transmission of all nerve signals. In white matter, $R1$ ('rate' of the '1st' process), the main DV in this study, is primarily driven by variation in myelin content.

- "**Retrogenesis** postulates that late maturing tissue is particularly vulnerable during aging and that tissue degeneration in the aging brain follows the reverse sequence of tissue maturation in the developing brain. This theory conceptualizes brain development like building a pyramid where the base is stabilized before additional layers are added. The top of the pyramid is the most vulnerable to aging-related decline, while the base remains sturdy. Retrogenesis **has not been formalized** in a manner that makes specific quantitative predictions, and several distinct hypotheses are discussed under the principle of retrogenesis."
- "Consistent with the retrogenesis hypothesis, in each fascicle **the rate of *R*1 development** as the brain approaches maturity **closely matches the rate of *R*1 degeneration** in aging."
- More formally, the **predicted** (by Jason et al) lifespan curve should be an **inverted-U-shaped curve that is symmetric about the maximum**.

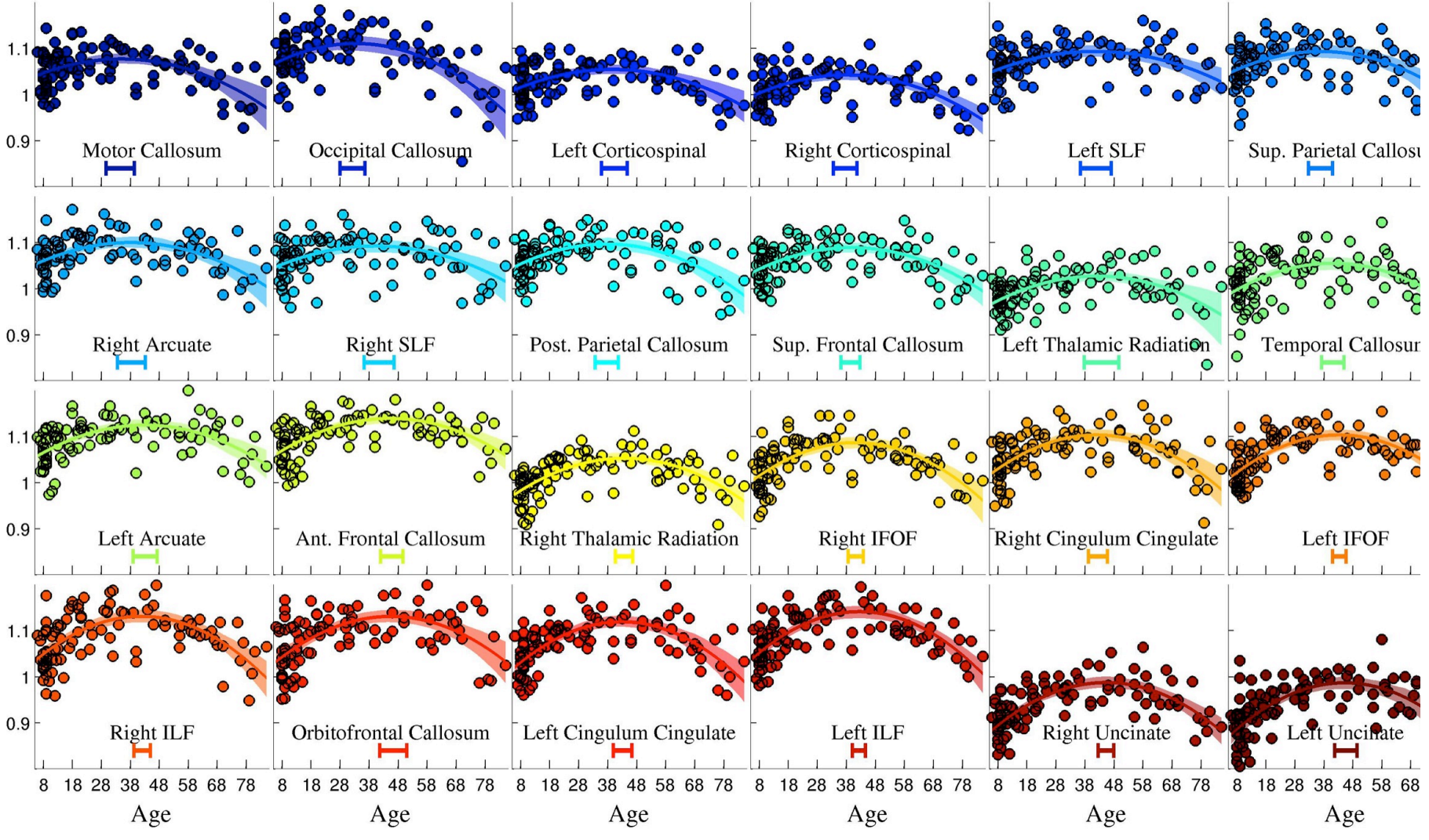
Jason's explanation of R1 (skip!)

- "In MRI you send energy into a solution (in our case a brain), wait some time, and then measure the energy that is emitted by the solution. Back in the 1940s it was noted that the rate at which any solution in a magnetic field loses energy is well characterized by exponential decay. Actually to be more precise there are two sets of equations that describe how two orthogonal components of the signal decay. The time constants in these two equations are referred to as T1 and T2. The reciprocal of T1 and T2 are R1 and R2. These time constants, in brain tissue, are highly dependent on myelin. R1 has nicer properties than T1, namely that it sums linearly when your measurement volume contains substances with differing R1 rates."




Lifespan change in R1 (1 s⁻¹)
0.020
0.045
0.090

Plots showing growth of tissue over the lifespan within each fascicle. A **second order polynomial model** fits the data as well than any more complex model. [How to handle **within-** and **between-**Ss differences?]



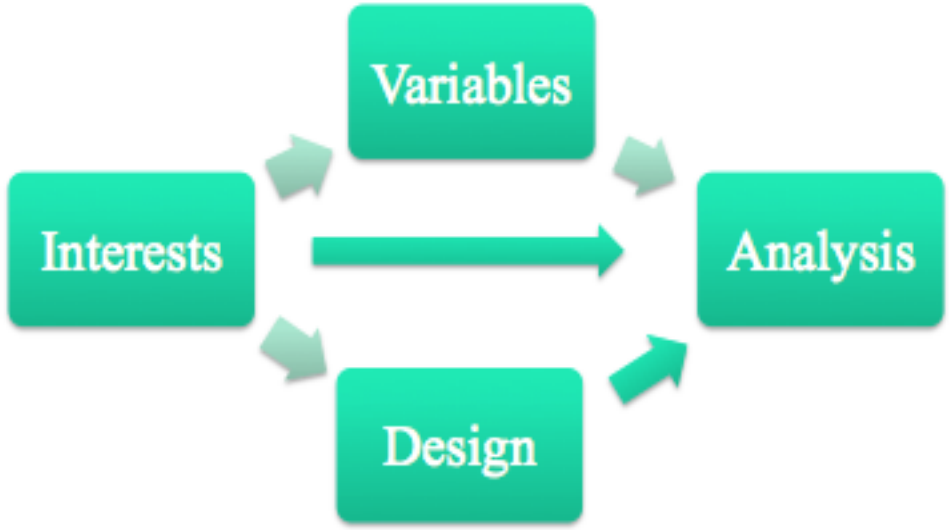
An approach to data analysis

(Illustrated in previous examples)

- Numerical data prompts the **curious** to ask: 'Is there a pattern, or general law?' Or, one's **theory** suggests the pattern to be confirmed in the data.
- **Plot** data for clues about pattern.
- *What is the **simplest**, interesting model? Does this model fit the data **significantly better** than the **null** model? What **statistical** model to use for **hypothesis test**? Check **model assns**.*
- What is a more complex model? Is it **sig. better** than the 'simplest' model? What **test** for this?

Lectures 1 & 2 outline

- **Class Project** on Memory Biases: Choice of statistical analysis is influenced by our *interests* (hypotheses we wish to test), the types of *variables* used (e.g., categ vs quant), and the *design* of our project (e.g., between- vs within-Subject).



Lectures 1 & 2 outline

- **Class Project**
- **Review**: types of variables, key concepts, *t*, χ^2 - appropriate terminology
- **Preview** of GLM: ANOVA, `lm()`, plots, interactions; key theoretical results; contrasts; causal diagrams
- **R**: examples

1. Class Project on memory biases

- Hand out questionnaires to class. Read instructions. Collect data.

- **Task**. Recall when you missed your plane or train, and then answer the questions on the questionnaire about how you felt, etc.
- **Data:** Collect *Pasthapp* scores in **R console**: 'free', xf = c(.., …); 'biased', xb = c(.., …); 'varied', xv = c(.., …)
- Describe published study. Is Class Project a good approx?

2. Summary of Morewedge *et al:* Hypotheses

Recall *an* instance [free]; or *the worst instance* [biased]; or *two or three instances* [varied] (This is a **between-subjects factor**) in which you missed your plane or train. Rate your feelings, etc.

The prediction of our **future emotional state**, if an event were to occur, is based on our **remembered state(s)** when similar events occurred in the past.

- H_1: correl(*Pasthapp, Futurehapp)* > 0.

When asked to recall a negative (positive) event, people tend **to remember extreme events**, i.e., events that are more negative (positive) than the typical event.

- H_2: *Pasthapp* [free] = *Pasthapp* [biased] < *Pasthapp* [varied],

Because the biased recallers were explicitly asked to recall 'the worst instance', they ought to **be aware that their level of *Pasthapp* is biased downwards**. Therefore, they should be able to correct for this bias when they predict their future happiness, *Futurehapp*.

- H_3: *Futurehapp* [free] < *Futurehapp* [biased] = *Futurehapp* [varied]

Morewedge *et al:* Results.1

- *Pasthapp* [free] = 23 (S.D. = *18*); *Pasthapp* [biased] = 20 (*27*); *Pasthapp* [varied] = 61 (*31*).

- The group differences are significant ($F(2, 59) = 16.0$, $p < .001$), as shown by a **1-way ANOVA**. **Planned orthogonal contrasts** are consistent with the authors' predictions.

Morewedge *et al:* Results.2

- *Futurehapp* [free] = 31 (*23*); *Futurehapp* [biased] = 46 (*26*); *Futurehapp* [varied] = 49 (*24*).

- The group differences are significant ($F(2, 59) = 3.29$, $p = .044$), as shown by a **1-way ANOVA**. **Planned orthogonal contrasts** are consistent with the authors' predictions.

Morewedge *et al:* Results.3

3. **Correlations**

Group (*n*)	***r*(Pasthapp, Futurehapp)**	**1-tailed *p*-value**
Free (19)	0.28	0.1228
Biased (19)	0.41	0.041

To get the *p*-values for *r*, I used the applet:

http://faculty.vassar.edu/lowry/tabs.html

3. Testable relationships in the Class Project

- **Additional Hypotheses**. What hypotheses might you entertain about the effects of *Responsible, Changes* and *FTP*?

Data from a previous sample

memgrp1	memgrp2	memgrp3
9.00	1.00	9.00
.00	5.00	8.00
3.00	7.00	5.00
5.00	2.00	17.00
5.00	.00	9.00
5.00	3.00	15.00
.	4.00	20.00

Possible analyses

- ***t*-tests** for testing if mean for ‘memgrp1’ is the same as, or different from, mean for ‘memgrp2’; also for ‘memgrp1’ vs. ‘memgrp3’, and ‘memgrp2’ vs. ‘memgrp3’. Do an example in R console:

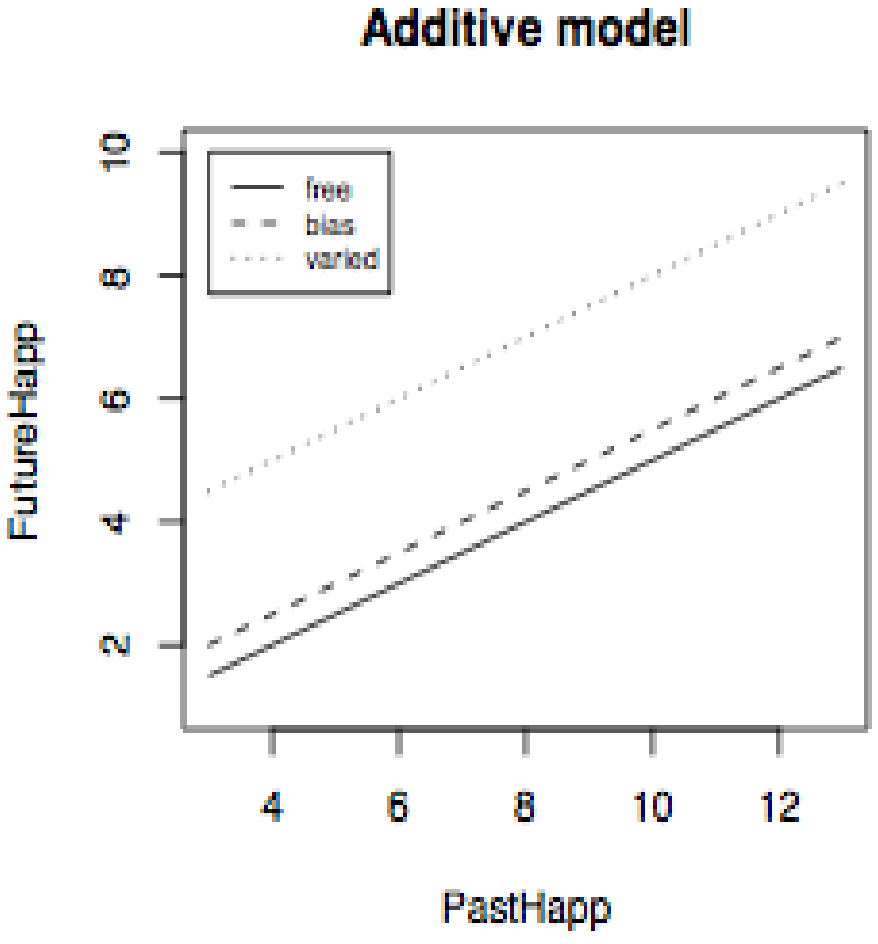
```
t.test(xf, xb, paired=F, var.equal=T,
na.action=na.omit)
```

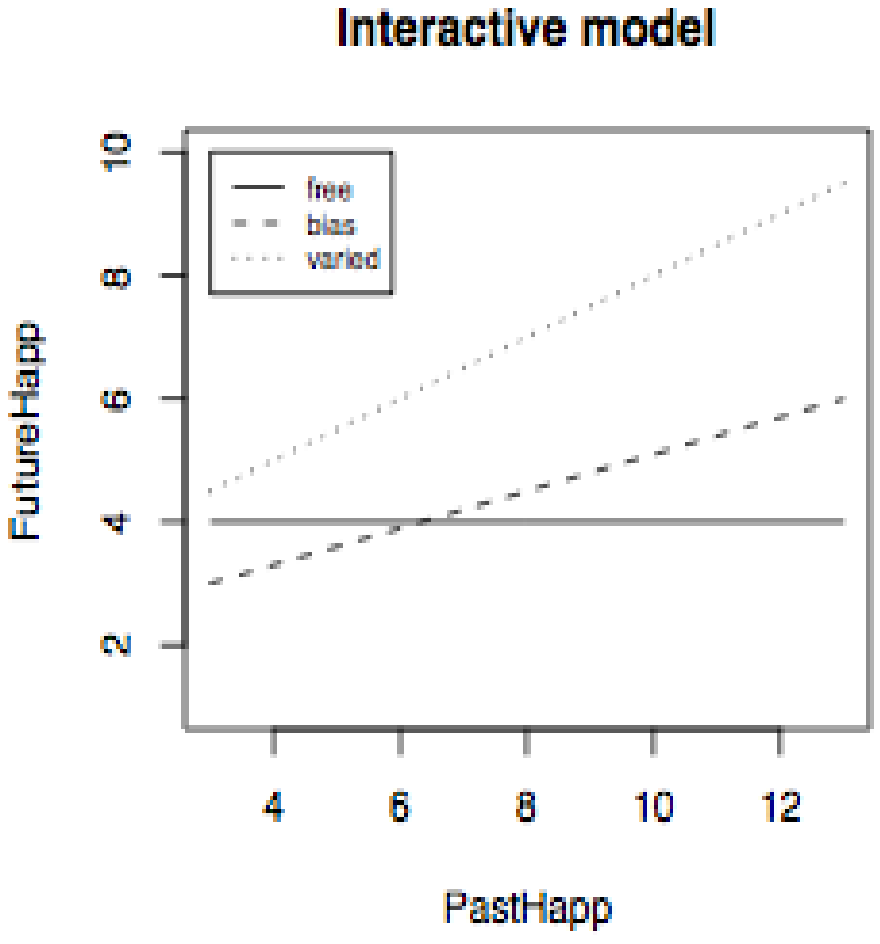
- **1-way ANOVA** and the associated ‘omnibus’ *F* test for testing if there is *some* difference among the 3 groups.

***F*-test for interaction** (more interesting!):

(a) Is the effect of the memory prime the same for ‘future-oriented’ people (*FTP* = ‘future’) as for ‘present-oriented’ people (*FTP* = ‘present’)?

(b) Is the relation between *Futurehapp* and *Pasthapp* the same for all 3 primes? If not, we say that there is an **interaction;** otherwise, memory prime and *Pasthapp* have **additive** effects on *Futurehapp*.


Additive model
free
bias
varied
FutureHapp
PastHapp
2
4
6
8
10
4
6
8
10
12


Interactive model
free
bias
varied
FutureHapp
PastHapp
2
4
6
8
10
4
6
8
10
12

Lectures 1 & 2 outline

- **Review**: types of variables, key concepts, *t*, χ^2 - appropriate terminology
- **Preview** of GLM: ANOVA, `lm()`, plots, interactions; key theoretical results; contrasts; causal diagrams
- **R**: examples

4. Some important ideas and distinctions

- **4.1.** *Sample* versus *population*

	Sample	**Population**
Descriptives	*Statistics*, e.g., $\bar{x}, s, \hat{p}, and\ r$	*Parameters*, e.g., mean, μ; s.d., σ ; probability, p; correlation, ρ.
Hypotheses (Inference)	**Not** $\bar{x}_1 = \bar{x}_2$	E.g., $\boldsymbol{\mu_1 = \mu_2}$

4.2. Variables

- .

	Qualitative vars.	**Quantitative vars.**
Values	*Changes* = Y or N *Memgrp* = 1, 2 or 3 *Ethnicity* = …	*Age* = 12.2 *FTP* = 12 *Pasthapp* = 7.3
Statistics	Frequency, f_i, Rel. Freq, rf_i; mode	Mean, s.d.; s.e.
Typical tests	χ^2 , Z	Z, t, F, r

4.3. Relationships

- The interesting questions involve a **relationship** between 2 (or more) variables, *X* and *Y*. For example, "Is mean(*Pasthapp*) equal for 'memgrp1' and 'memgrp2'?" concerns a **difference: "Is $\mu_1 = \mu_2$?"**.
- However, this question can be restated as, "Is there a **relationship** between 'memory group' (*X)* and *Pasthapp* (*Y).*" So *t*- and *F*-tests for differences can be redone using correlation and regression techniques (which we use to test for relationships).

4.3.1. A common question at the intersection of 'difference' & 'relationship'

De: Ben Bolker <bbolker@gmail.com>

Para: r-sig-mixed-models@r-project.org

Enviado: sábado 24 de septiembre de 2011 23:02

Asunto: Re: [R-sig-ME] Factor collapsing method

Iker Vaquero Alba <karraspito@...> writes:

… I get a significant effect in a factor with 7 levels … but I can't know which of the levels are the most important ones in determining that effect. I have an idea from the p-values of the "summary" table, and I can also plot the data to see the direction of the effect. However, I have read in a paper that there is **a method to collapse factor levels to obtain information about which factor levels differ from one another**, that is used when an explanatory variable has a significant effect in the minimal model and contains more than two factor levels. I have looked for it in the Crawley book and in the web, but I actually cannot find anything ...

Discuss *post hoc* tests! `rs1 = lm(y ~ A); plot(TukeyHSD(rs1))`

Lecture 2 outline

- **Toggle** among (i) terminology & methods (t, χ^2), (ii) advanced methods preview ('interaction', `lm()`), and (iii) basic theorems – in a 'nonlinear' way!
- **Review** meaning of the **interaction** between X_1 and X_2 in their effects on Y. Use `lm()`.
- **Review** possible analyses on 'memgrp' data; related theory.
- **Monte Carlo** methods: CLT states that the distrn of a sum is approximately Normal. For a given non-Normal popn distrn, how good is this approx?
- **Preview**: Introduce 'fieldsimul1.csv' to preview GLM. (*Apply to your own data now!*)

4.4. Causal Models in Flowcharts

If X and Y are 2 variables in a data set, it is possible that

(a) X causes Y, $\boldsymbol{X \rightarrow Y}$;

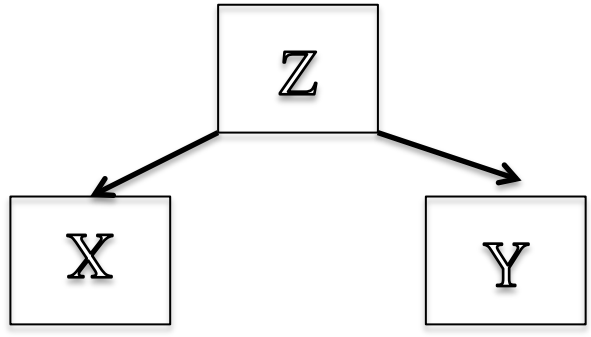
(b) Y causes X, $\boldsymbol{Y \rightarrow X}$;

(c) both (a) and (b), $\boldsymbol{X \leftrightarrow Y}$;

(d) neither (a) nor (b), $\boldsymbol{X \quad Y}$;

(e) Z causes X and Y: $\boldsymbol{Z \rightarrow X}$ and $\boldsymbol{Z \rightarrow Y}$ (spurious);

(f) $\boldsymbol{X \rightarrow Z \rightarrow Y}$ (mediation).



4.4.1. Mediators

- Consider the SES of our parents (PSES), our own SES, and our level of education (Educ); with Educ as a **mediator** variable:
- ***PSES → Educ → SES***

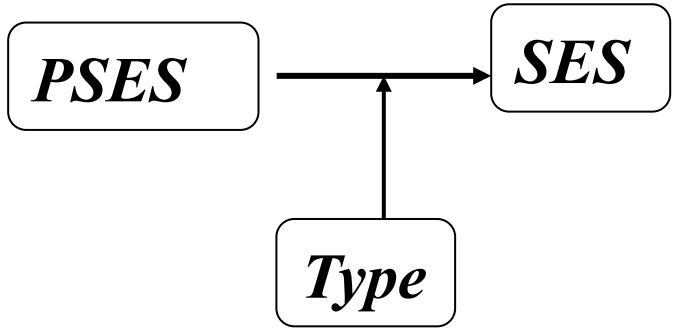
4.4.2. Moderators

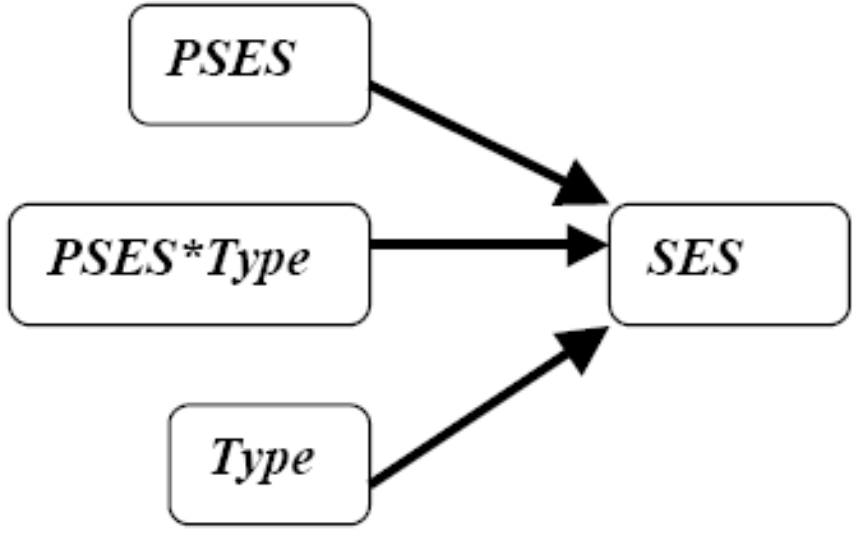
- Consider PSES, our own SES, and the Type (e.g., MDC vs. LDC) of society. Suppose
  - cor(PSES, SES) > 0 in Type A societies,
  - cor(PSES, SES) = 0 in Type B societies.
  - i.e., PSES is uncorrelated with SES in Type B, but not Type A, societies.
- Here, Type is a **moderator** variable; it moderates the effect of PSES on SES. Also, there is an **interaction** between PSES and Type in their effects on SES.

Diagrams for Moderation

(Useful tool in setting out one's causal theory!)

- .





4.5. Between- and Within-subjects research designs

- ‘Memory group’ is a **between-subjects** factor in the class project.
- In contrast, we could have used a **within-subjects** design in which each subject is exposed to all 3 levels of the factor. Then, each level of the factor would be associated with the same group of subjects. We would then compare the effects of the different primes **within** each participant.

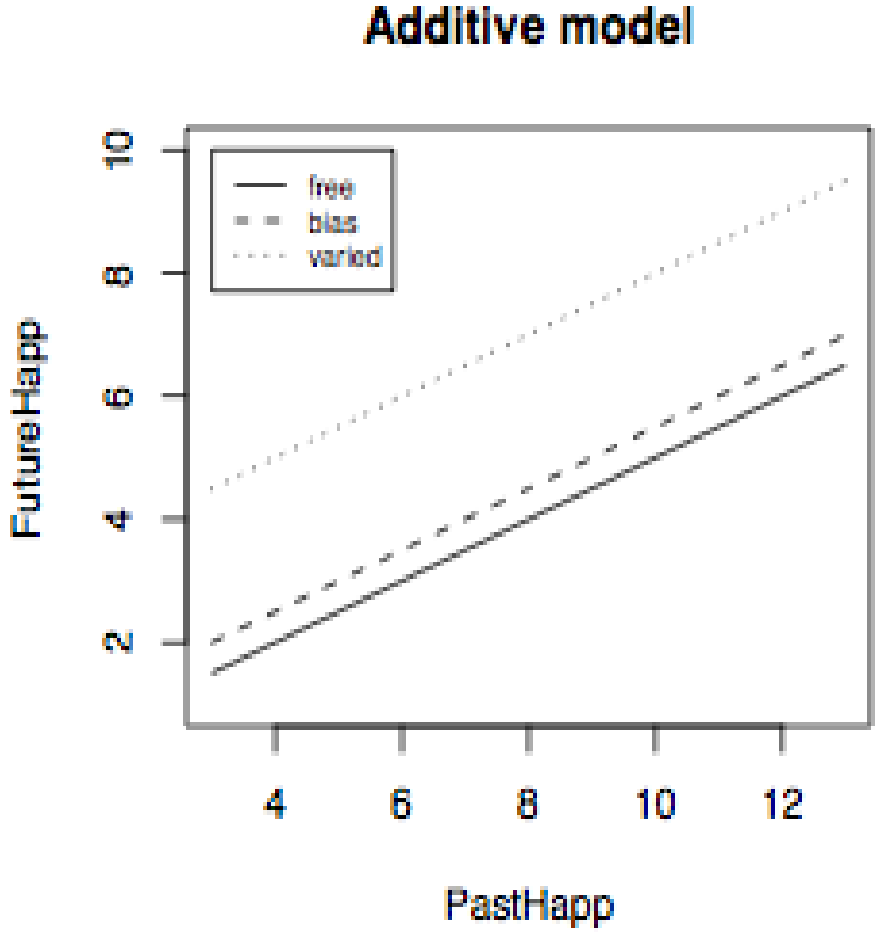
- Pros and Cons? Would *you* use a within-subject design to study 'memory group' effects? Why not?

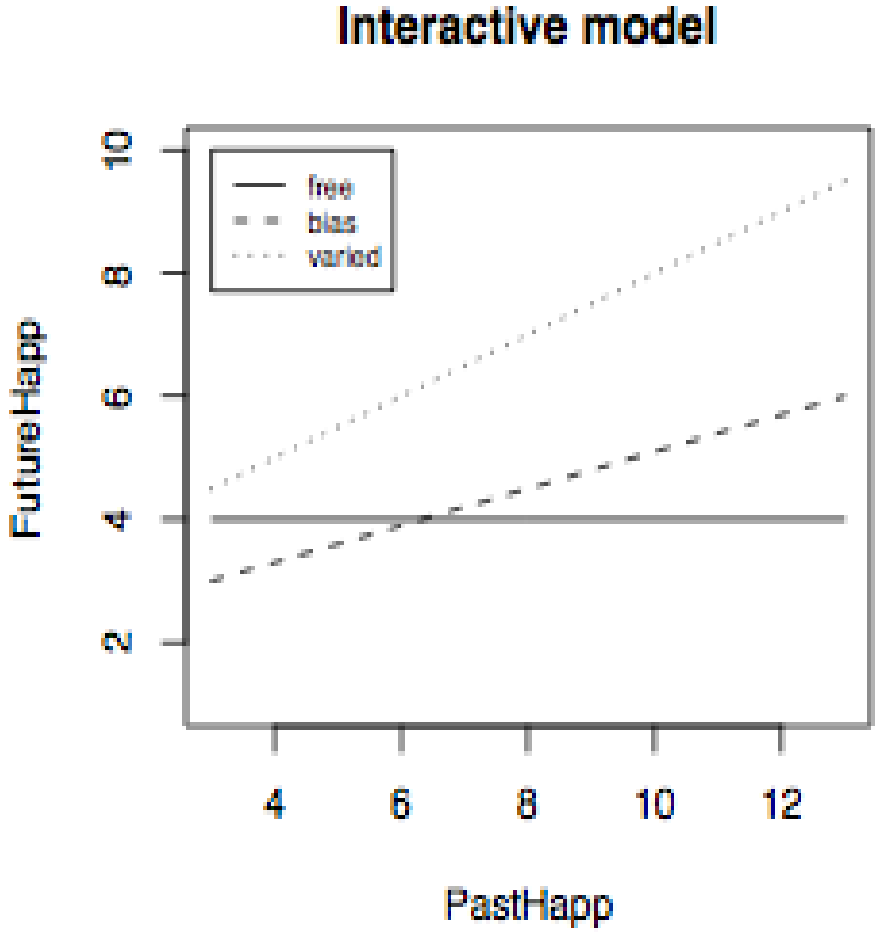
- **The appropriate statistical analysis is different for between-subjects designs and within-subjects designs.** Designs with both between-subjects and within-subjects factors are called **mixed designs**.

Measuring the interaction between 'mem grp' (X_1) and *Pasthapp* (X_2) on *Futurehapp* (*Y*)

- Consider two possible ways (models) in which 'memory group' and *Pasthapp* can jointly affect *Futurehapp*
- **Additive model**: differences in *Futurehapp* among the memory groups are approximately the same at all levels of *Pasthapp*.
  - 'memory group' and *Pasthapp* do not **interact**; they have additive effects on *Futurehapp*, and the curves are **parallel**. This is the critical visual (but informal) test for the absence of an interaction.

Measuring the interaction between ‘mem grp’ (X_1) and *Pasthapp* (X_2) on *Futurehapp* (Y)





- **Interactive model**: the differences in *Futurehapp* among the memory groups depend on the level of *Pasthapp*.
  - 'memory group' and *Pasthapp* do **interact** in their effects on *Futurehapp*. The curves are **not parallel** – this is the critical visual (but informal) test for the presence of an interaction.
- A formal statistical test would test for the presence of an interaction (or the absence of additivity, or the non-parallelism of the curves) by means of an ***F*-, or *t*-test**.

GLM with lm()

1-way ANOVA

```
rs2 = lm(phapp ~ memgrp, na.action=na.omit, d0)
print(summary(rs2))
```

GLM with 1 categorical predictor ('memgrp') and 1 quantitative predictor ('phapp'); DV is 'future happ' ('fhapp')

```
rs3 = lm(fhapp ~ phapp + memgrp, na.action=na.omit,
d0)print(summary(rs3))          #additive model
rs3a = lm(fhapp ~ phapp * memgrp, na.action=na.omit, d0)
## rs3a is SAME as rs3b
rs3b = lm(fhapp ~ phapp + memgrp + phapp : memgrp,
na.action=na.omit, d0)          #interactive model
print(summary(rs3a))
print(anova(rs3, rs3a))         #model comparison
```

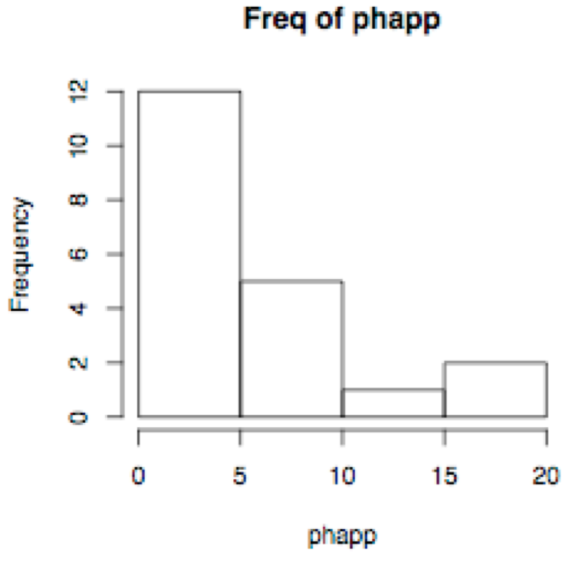
```
lm(formula = Futurehapp ~ Pasthapp + memtype, data =
   dat0, na.action = na.omit)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.5851     1.1350   2.278   0.0280 *
Pasthapp       0.3372     0.1406   2.399   0.0211 *
memtypebias    0.3488     1.3500   0.258   0.7974
memtypevaried  0.1541     1.3497   0.114   0.9096
---
Residual standard error: 3.592 on 41 degrees of
   freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.1462,    Adjusted R-squared:
   0.08373
F-statistic:  2.34 on 3 and 41 DF,  p-value: 0.08742
```

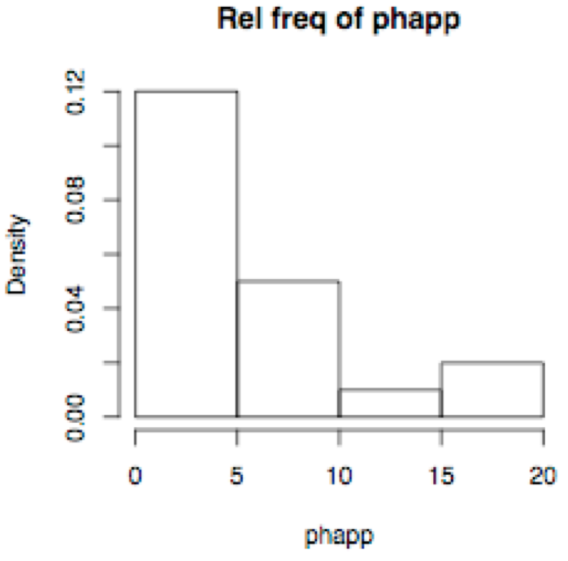
```
lm(formula = Futurehapp ~ Pasthapp * memtype, data =
   dat0, na.action = na.omit)

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.87262    1.04607   2.746  0.00908 **
Pasthapp           0.28684    0.17455   1.643  0.10836
memtype1           0.33030    1.31053   0.252  0.80234
memtype2          -0.33671    0.72234  -0.466  0.64371
Pasthapp:memtype1 -0.04898    0.24611  -0.199  0.84328
Pasthapp:memtype2  0.05816    0.10137   0.574  0.56947
---
Residual standard error: 3.663 on 39 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.1553, Adjusted R-squared: 0.04705
F-statistic: 1.435 on 5 and 39 DF,  p-value: 0.2336
```

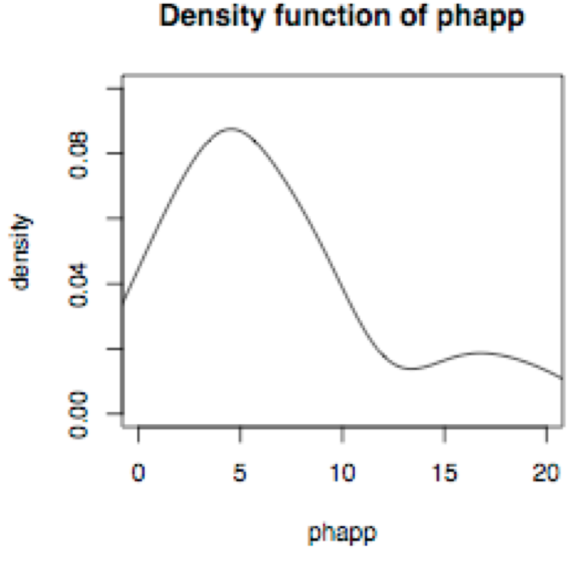
Statistics from ‘Mem Bias’ project

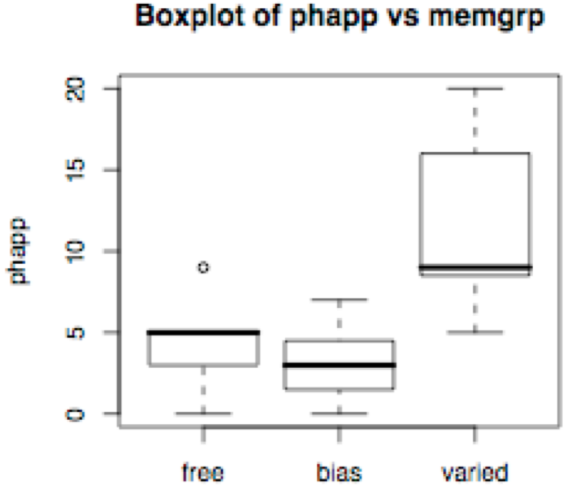
- Freq distrn, Relative freq distrn, Probability Density Function (pdf)
- Boxplot: mean & variability in ‘phapp’ at each level of ‘memgrp’
- T-test: difference between 2 means
- 1-way ANOVA, using lm(), the workhorse GLM function in R

Distributions









t-tests

- Sample of a random variable, X: $x_1, x_2, \ldots, x_n$; from which we calculate the sample mean, $\bar{x}$, st. dev., s, and variance, s^2.
- Z-scores (or standard scores), defined as $z_i = (x_i - \bar{x})/s$. **The mean of all *z*-scores is 0, and the s.d. is 1.** 'Large' values of z are in the ranges **'±2 or more extreme.'**
- **Linear transformations**: $Y = a + bX$.
  - Mean: $\mu_Y = a + b\,\mu_X$
  - S.d.: $\sigma_Y = b\sigma_X$. Variance: $\sigma_Y^2 = b^2\sigma_X^2$

If X has a (parent) distrn, N(μ, σ^2), i.e., Normal with mean, μ, and s.d., σ, then we can convert the sample mean, $\bar{X}$, to a z-score or a t-score:

$$Z = \frac{Variable - (Its\ Mean)}{(Its\ SD)}$$

$$= \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}.$$

$$t_{n-1} = \frac{\bar{X} - \mu}{s / \sqrt{n}}, \text{ with } n\text{-1 df.}$$

t versus *Z*

- *E(Z) = 0; E(t_k) = 0 (k* is the df of *t)*.
- *Var(Z) = 1*

$$\text{var}(t_k) = \frac{k}{k-2}, k > 2;$$

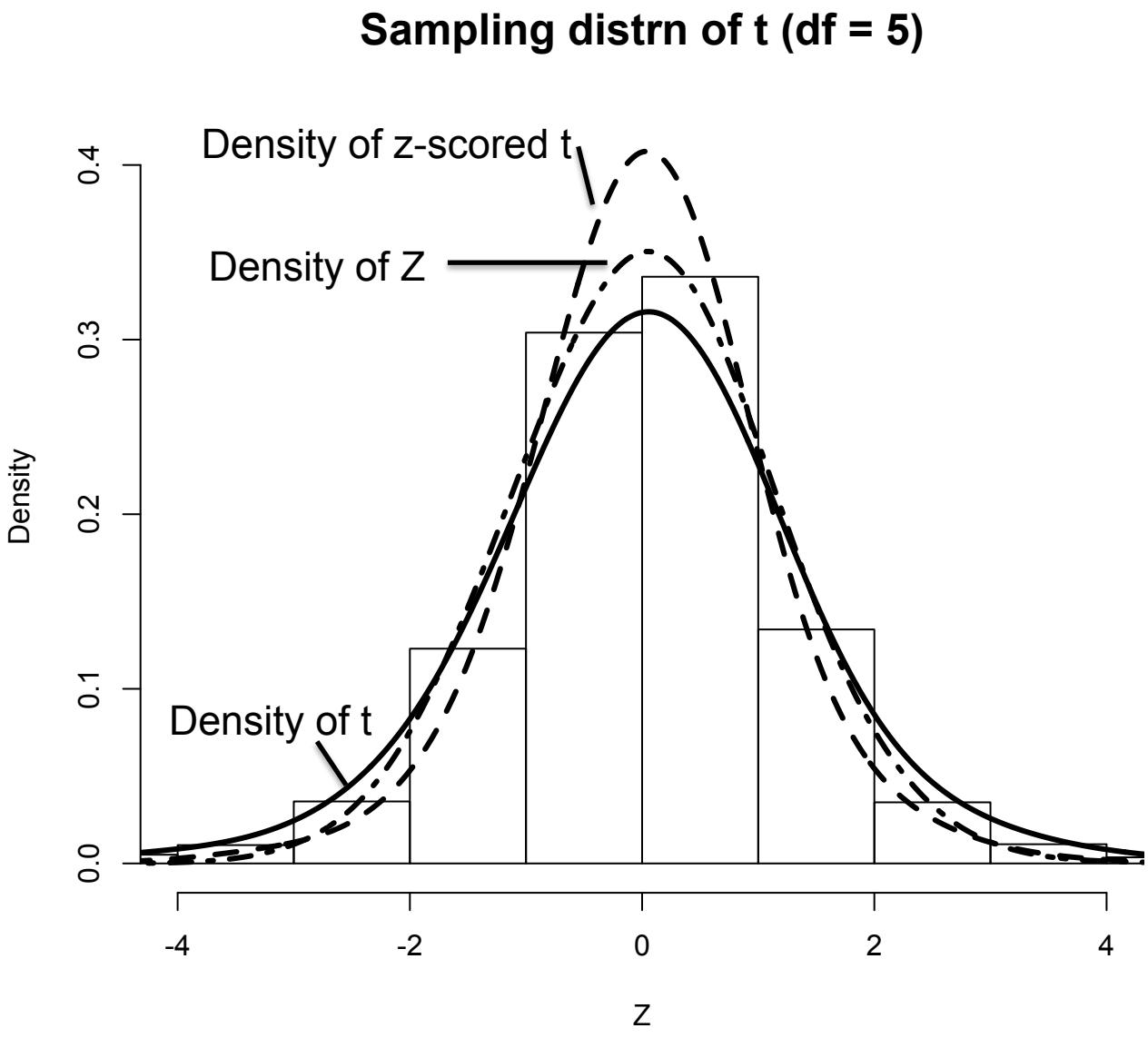
$$sd(t_k) = \sqrt{\frac{k}{k-2}}, k > 2;$$

$$Z = \frac{t_k}{\sqrt{k/(k-2)}}$$

A comparison of the *t*- and *Z*-distrns

```
z0 = rnorm(1000)  # sample from N(0,1)
t0 = rt(1000, df = 5) # sample from t(5)
zt0 = t0*(3/5)^0.5   # t_k/(sd(t_k))
```

1. Plot histogram of t0 and density of t0 (solid curve); note good approx.
2. Compare density of t0 with densities of z0 & zt0 (dashed curves); note t0 has greater sd ((5/3)^.5 = 1.29) than z0 or zt0 (sd = 1).
3. The density of zt0 is MORE peaked than that of z0 (the Normal) - *kurtosis*.

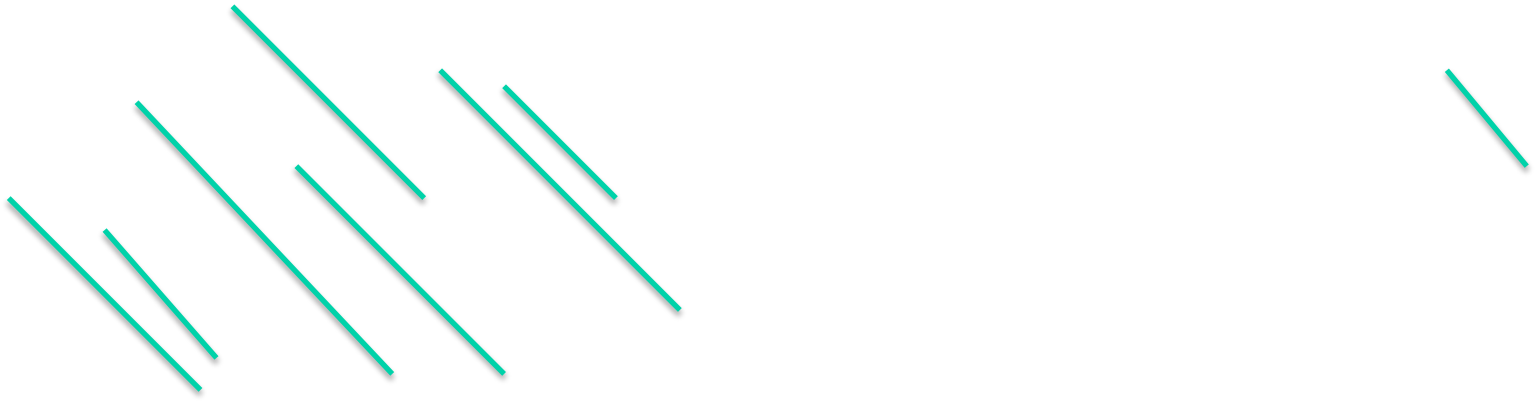

Sampling distrn of t (df = 5)
Density of z-scored t
Density of Z
Density of t
Density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
Z

Lecture 3 outline

- **Decision-making under uncertainty:** the role of variability and amount of data.
- **Review** of the chi-square (χ^2) goodness-of-fit test, and the chi-square (χ^2) contingency test, using **'fieldsimul1.csv'**.
- Also **preview** GLM

Categorisation.1

Observed exemplars of X's **Test**



- Is the Test line a member of Category X: Y or N?
- How **confident** are you?
- *Generalisation gradient*: If *Y* is your typical response to X's, what is your response, *y(Test, X)*, to Test?

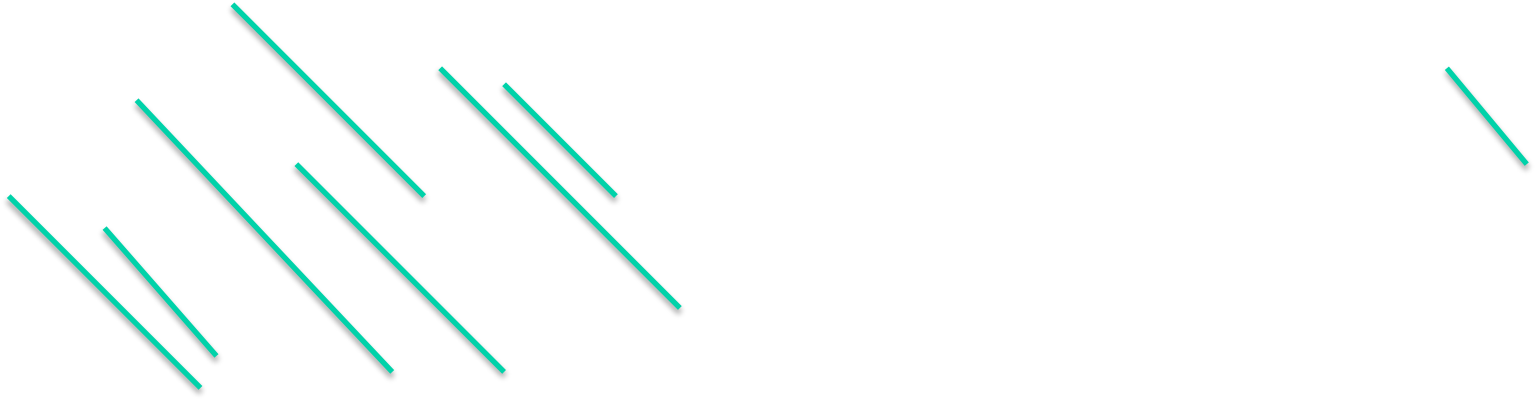
Categorisation.2

Observed exemplars of X's **Test**

- Is the Test line a member of Category X: Y or N?
- How **confident** are you?

Categorisation.3

Observed exemplars of X's **Mean of 20 Tests**



- Did the *n* Tests come from Category X: Y or N?
- How **confident** are you?
- **What factors** influence categorical judgments?

Role of *variance*

- Let $x_1, x_2, \ldots, x_n$ be the line lengths of the observed exemplars; and z is that of the test. The Sum of Squares, SS, and the variance, s^2, of the $\{x_i\}$ are given by:

$$\text{Sum of Squares, } SS = \sum_{i=1}^{n} (x_i - \overline{x})^2,$$

$$\text{and the variance, } s^2 = \frac{SS}{(n-1)}.$$

- If the variance is 'large' (for a fixed mean), then z is more likely to be seen as an X than if the variance is 'small'. This is why 'Test' is more likely to be judged as an X in Categorisation.1 than in Categorization.2.

Role of *similarity*

- Michael's model made use of the important concept of **similarity** between z and each x_i. The 'similarity' between z and the $\{x_i\}$ is less in Categorisation.2 than in Categorization.1. But we have to define 'similarity'.
- I asserted in class today that 'similarity' is a more complex concept than 'variance', but that it could be related to it. Here is one way of doing so.
- It is reasonable to **assume** that the 'similarity' between z and the $\{x_i\}$ is **inversely related** to the deviation, $z - \overline{x}$, between z and the mean, **after this deviation is standardised**; i.e., inversely related to $(z - \overline{x}) / s$.

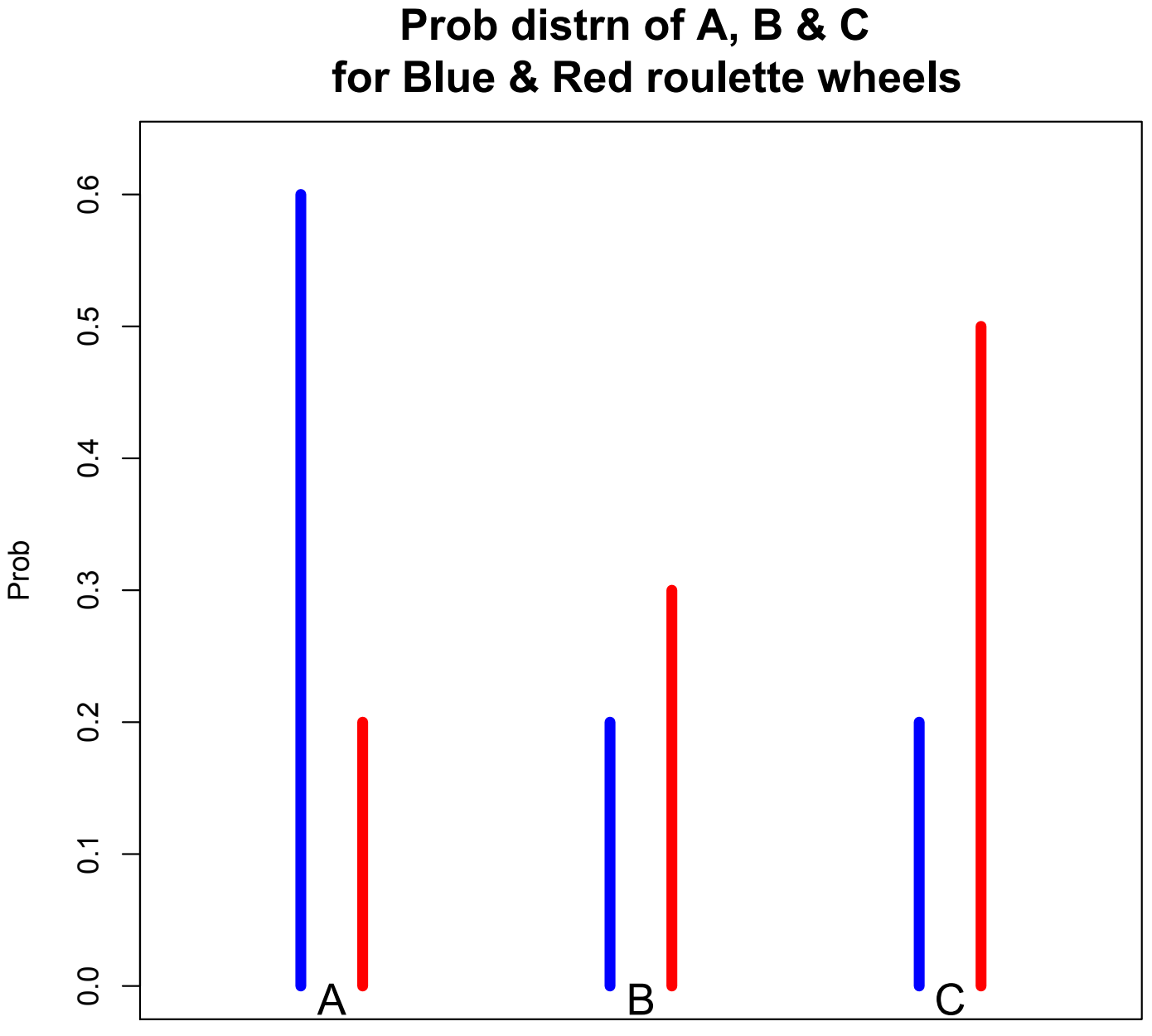
Role of *similarity*

- With this **assumption** that the 'similarity' between z and the $\{x_i\}$ is **inversely related** to
- $$(z - \overline{x}) \Big/ s$$:
- We can see that 'similarity' is **directly** related to s (for a given mean and a given test) and, therefore, to variance; and that, therefore, 'similarity' is higher in Categorisation.1 than Categorisation.2.

Categorisation.4

- On each trial in an expt, one of two stimulus types (e.g., signal vs. noise, Categ 1 vs. Categ 2, Wheel 1 vs Wheel 2) is chosen and a stimulus is generated.
- S's task is to decide which 'type' was used to generate the stimulus on that trial.
- Here, we have a Blue or a Red roulette wheel. Each generates 3 possible stimuli, A, B or C, with different known probabilities.
- If an A is observed, what would be your response? Or a B? Or a C?
- What **'principle' or 'rule'** are you using?

Categorisation.4



Obs. **Source?**

A ____

B ____

C ____

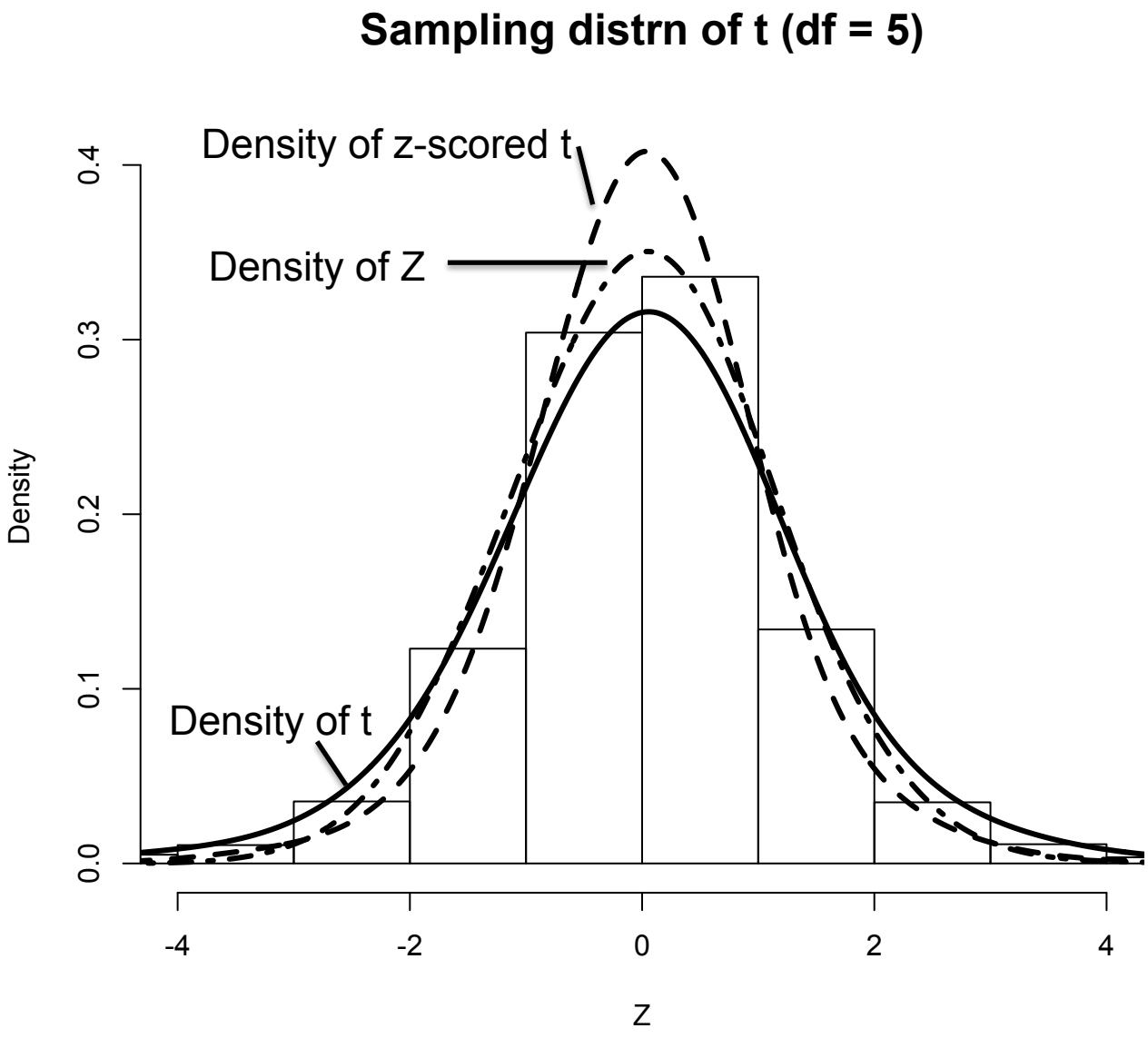
Some useful concepts

- Variance of a popn, var(X).
- Variance of a statistic, e.g., a sample mean that is based on n observations.
- The probability or likelihood of an observation, and the **principle of maximum likelihood**.
- Inference requires us to know not only what was observed, but also what **might have been** observed but wasn't.

Illustration of the *t*- and *Z*-distrns

```
z0 = rnorm(1000)  # sample from N(0,1)
t0 = rt(1000, df = 5) # sample from t(5)
zt0 = t0*(3/5)^0.5   # t_k/(sd(t_k))
```

1. Plot histogram of t0 and density of t0 (solid curve); note good approx.
2. Compare density of t0 with densities of z0 & zt0 (dashed curves); note t0 has greater sd ((5/3)^.5 = 1.29) than z0 or zt0 (sd = 1).
3. The density of zt0 is MORE peaked than that of z0 (the Normal) - *kurtosis*.


Sampling distrn of t (df = 5)
Density of z-scored t
Density of Z
Density of t
Density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
Z

If X has a (parent) distrn, N(μ, σ^2), i.e., Normal with mean, μ, and s.d., σ, then we can convert the sample mean, $\bar{X}$, to a z-score or a t-score:

$$Z = \frac{Variable - (Its\ Mean)}{(Its\ SD)}$$

$$= \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}.$$

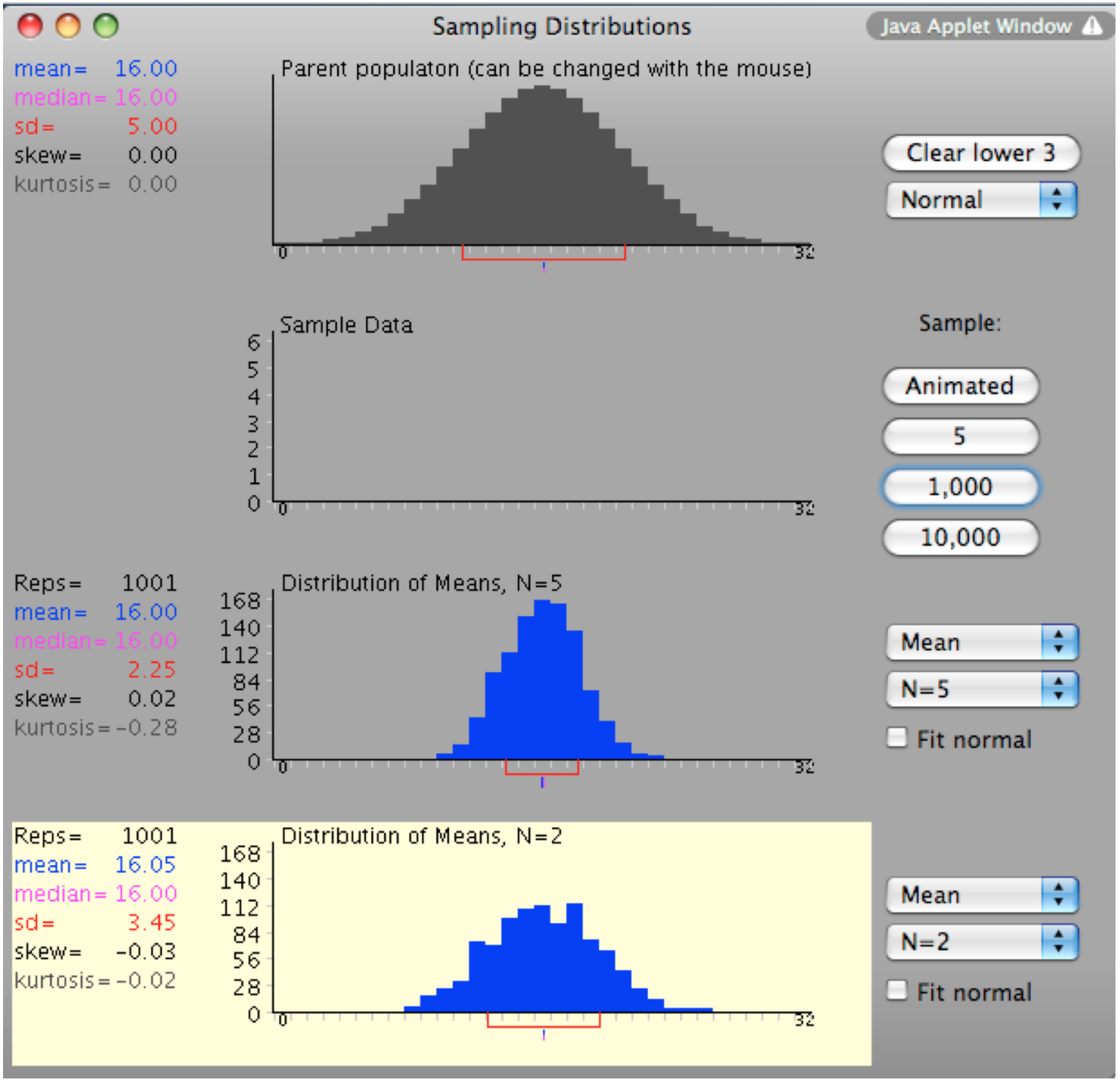
$$t_{n-1} = \frac{\bar{X} - \mu}{s / \sqrt{n}}, \text{ with } n\text{-1 df.}$$

The Central Limit Theorem (CLT)

- Under certain mild conditions on the population distrn of X, the distribution of the sample mean, $\overline{X}$, tends to the Normal distrn. If we convert the sample mean into a z-score, Z, then
- $Z \sim N(0, 1)$, i.e., Z is approximately a *standard* Normal random variable.
- If we do not know σ, then we convert the sample mean into a t score.

Sampling distributions of $\overline{X}$, and s^2, for different sample sizes, n

The applet on "Sampling Distributions" at http://onlinestatbook.com/stat_sim/sampling_dist/index.html illustrates various sampling distrns, and the 'effect' of the Central Limit Theorem. **Please play around with this applet.**


Sampling Distributions
Java Applet Window
mean= 16.00
median= 16.00
sd= 5.00
skew= 0.00
kurtosis= 0.00
Parent populaton (can be changed with the mouse)
0
32
Clear lower 3
Normal
Sample Data
6
5
4
3
2
1
0
0
32
Sample:
Animated
5
1,000
10,000
Reps= 1001
mean= 16.00
median= 16.00
sd= 2.25
skew= 0.02
kurtosis= -0.28
Distribution of Means, N=5
168
140
112
84
56
28
0
0
32
Mean
N=5
Fit normal
Reps= 1001
mean= 16.05
median= 16.00
sd= 3.45
skew= -0.03
kurtosis= -0.02
Distribution of Means, N=2
168
140
112
84
56
28
0
0
32
Mean
N=2
Fit normal

t-test for 2 independent samples

- The null is H_0: $\mu_1 = \mu_2$.
- The alternative is H_1: $\mu_1 \neq \mu_2$ (2-tailed test)
- The test statistic, *t*, is defined as follows:
  - The numerator of the *t*-ratio is the difference between the 2 sample means
  - The denominator is the **estimate** of the standard dev (also called 'standard error') of the difference between the 2 means
  - The degrees of freedom (df) of *t* is n_1+n_2-2.

t-test in R

- Suppose the data from 2 samples are the vectors, vf and vb.

```
rs1 = t.test(vf, vb, paired = F, var.equal = T, na.action = na.omit)
print(rs1)
```

- If the 2 samples were paired (and, therefore, not independent), we would use "`paired = T`".
- There is a test of 'homogeneity of variance', i.e., whether "`var.equal=T`"

Preferred data format (long form)

phapp	memgrp
9	1
0	1
3	1
5	1
5	1
5	1
1	2
5	2
7	2
2	2
0	2
3	2
4	2
9	3
8	3
5	3
17	3
9	3
15	3
20	3

e.g.:
1 = ‘free’
2 = ‘biased’
3 = ‘varied’

- `rs1 = t.test(phapp[memgrp==1], phapp[memgrp==2], paired=F, var.equal=T, na.action=na.omit)`

```
t = 0.914, df = 11, p-value = 0.3803
alternative hypothesis: true difference in
  means is not equal to 0
95 percent confidence interval:
 -1.911046  4.625331
sample estimates:
mean of x mean of y
 4.500000  3.142857
```

F-test for homogeneity of variance

- $H_0 : \sigma_1^2 = \sigma_2^2$.
- $H_1 : \sigma_1^2 \neq \sigma_2^2$.
- Let s_1^2 and s_2^2 be the variances of the 2 samples. Then the F-ratio for testing H_0 is
  - $F = \max\{s_1^2, s_2^2\}/\min\{s_1^2, s_2^2\}$, and 'large' values of F (e.g., $F > 4$) suggest that the null should be rejected.
  - Because of the way F is defined, a 1-tailed test is appropriate.

F-test in R

```
rs1a = var.test(vf, vb,
na.action=na.omit)
print(rs1a)

F test to compare 2 variances
F = 1.5, num df = 5, denom df = 6, p-
  value = 0.63
95 percent confidence interval:
0.25  10.4.
```

This CI contains 1; therefore, we cannot reject H_0.

Definition of mean, var, s.d.

- **Popn** distrn (discrete) with possible values, $x_1, x_2, \ldots,$ and associated probs, $p_1, p_2, \ldots,$ yields mean & variance:

$$\mu = \sum p_i x_i ; \sigma^2 = \sum p_i (x_i - \mu)^2 ; \sigma = \sqrt{\sigma^2}$$

```
d0 = c(0:6)   # x1=0, x2=1, …, x7=6
p0 = c(.1,.25,.45,.09,.07,.03,.01) # p1=.1, …
mu0 = sum(d0*p0)/sum(p0) # Mean or Expected Value of X
var0 = sum(p0*(d0 - mu0)^2)/sum(p0)# variance
sd0 = var0^.5 # s.d.
skw0 = sum(p0*(d0 - mu0)^3)/sum(p0)# skewness
print(c(mean=mu0, sd = sd0, skew = skw0))
```

Appendix

- *The remaining slides in this file contain theoretical results about sampling distributions. They will not be part of Lecture 3, and we hope to return to this material at a later date.*

The sum, *T*, of *n independent* observations of *X (HO-1*, pp 12-15*)*

- $T = \sum_{i=1}^{n} X_i.$

- $[Ans. \mu_T = n\mu_X;$

$$\sigma_T^2 \equiv \text{var}(T) = \sum_{i=1}^{n} \text{var}(X_i) = n\sigma_X^2;$$

$$\sigma_T = \sqrt{n}\sigma_X .]$$

The *difference*, *D*, between two independent variables

Recall the useful results on **linear transformations,** $Y = a + bX$.

Mean: $\mu_Y = a + b\,\mu_X$

S.d.: $\sigma_Y = b\sigma_X$. Variance: $\sigma_Y^2 = b^2\sigma_X^2$

$$\sigma_D^2 \equiv \mathrm{var}(D) = \mathrm{var}(X_1 + (-1 * X_2)) =$$

$$\mathrm{var}(X_1) + \mathrm{var}((-1) * X_2) = \mathrm{var}(X_1) + (-1)^2\,\mathrm{var}(X_2)$$

$$= \mathrm{var}(X_1) + \mathrm{var}(X_2)$$

The *mean*, $\bar{X}$, of *n independent* observations

$$\bar{X} = \left(\frac{1}{n}\right)\sum_{i=1}^{n} X_i = \left(\frac{1}{n}\right) T \ .$$

$$\sigma^2_{\bar{X}} \equiv \mathrm{var}(\bar{X}) = \left(\frac{1}{n}\right)^2 \mathrm{var}(T) =$$

$$\left(\frac{1}{n}\right)^2 (n\sigma^2_X) = \frac{\sigma^2_X}{n};$$

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$$