Lecture 1.3

Summary of Lec 1 & SG's Lec 2 (Thanks!!)

- Examples of qualitative research: values, thoughts, stereotypes, pretend play
- For categorical responses use Cohen's κ: to get *reliability* as agreement corrected for chance
- For count or other quantitative data, use Cronbach's α, or ICC, to measure *reliability* as *internal consistency* among items
- Attend to *unit of analysis*; and *data format*, e.g., is 'rater' a fixed or random effects factor?

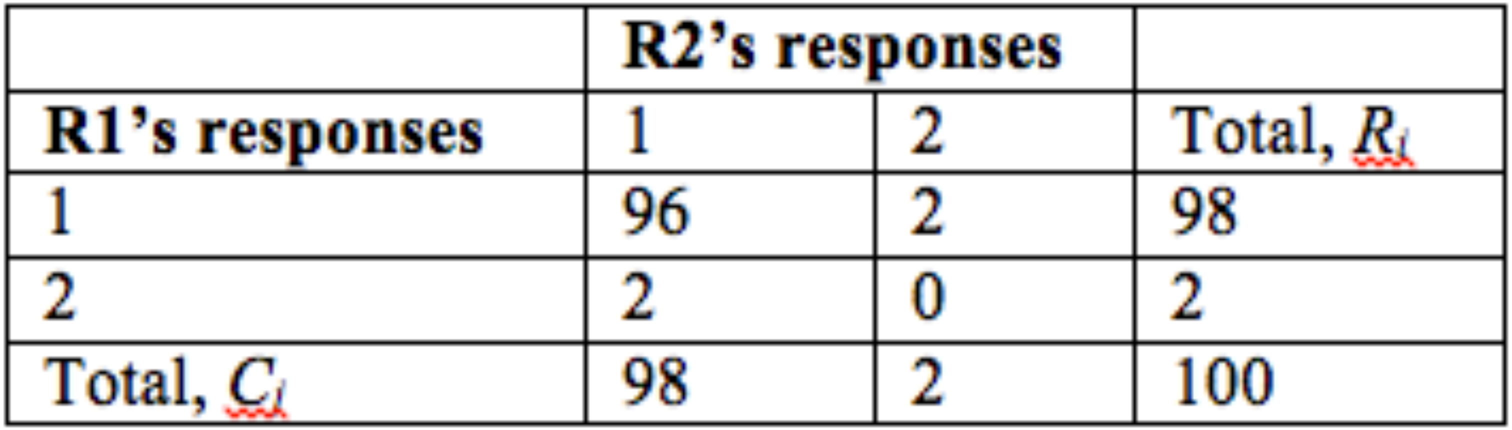
Outline of Lecture 1.3

- A problem with *κ* when a category is rare
- Cognitive Models offer a solution, but are also useful in their own right
- Maximum Likelihood (ML) estimation of model parameters; the bootstrap
- Bayesian estimation and MCMC
- Steve S's problem, when an 'expert' and a 'novice' are the 2 raters
- Pivot to theory-based approaches to Reliability

A problem with Cohen's *κ*,

when a category is rare

- For these data, agreement is 'high', but ***κ*** = -.02, very low! (It is as if raters were told to find the two "2"s in the set of 100, and they do not agree on either of the two.)
- *κ* is weak conceptually, inasmuch as there is no cognitive process to which it corresponds.

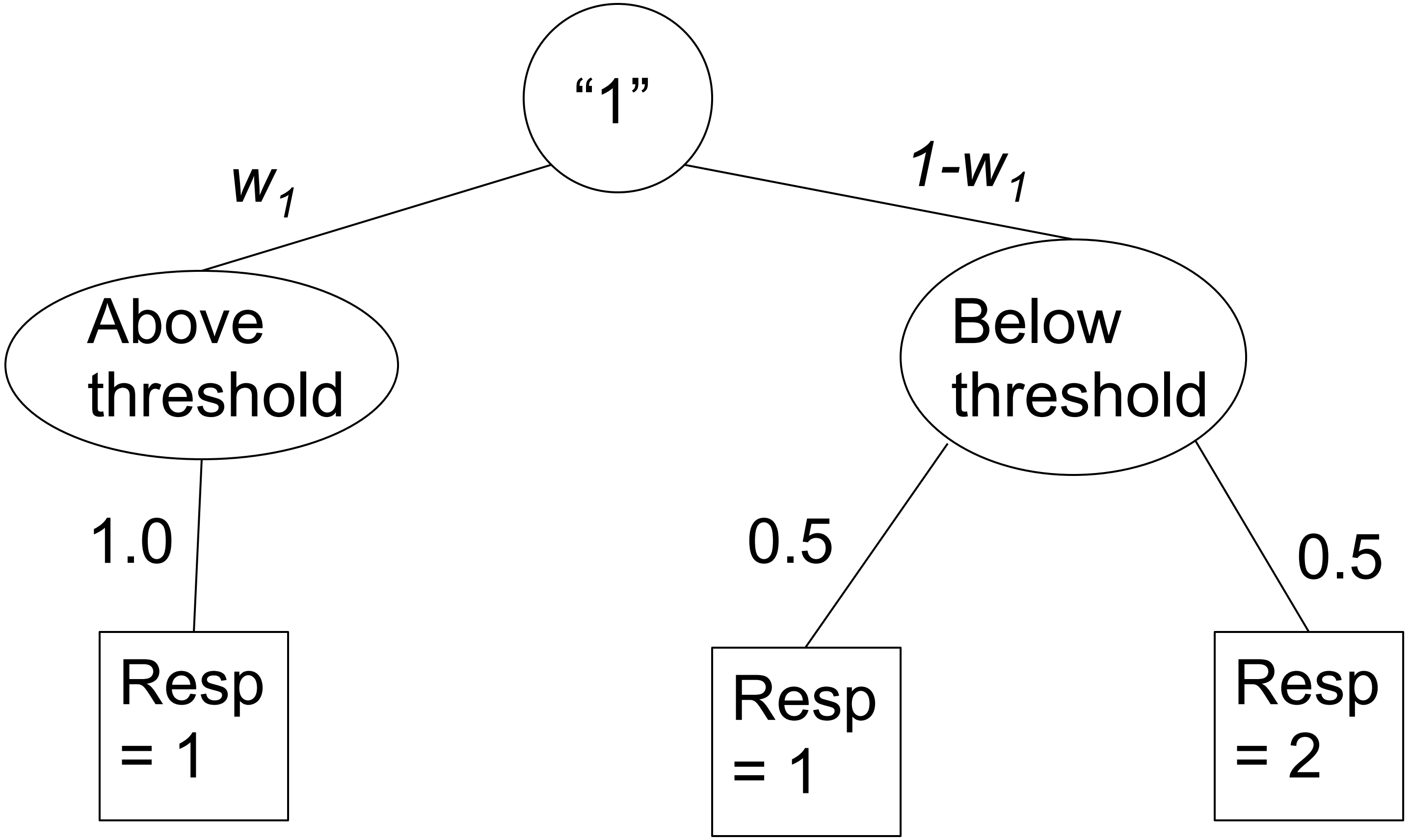


	R2's responses		
R1's responses	1	2	Total, R_i
1	96	2	98
2	2	0	2
Total, C_i	98	2	100

A solution via a Cognitive Model

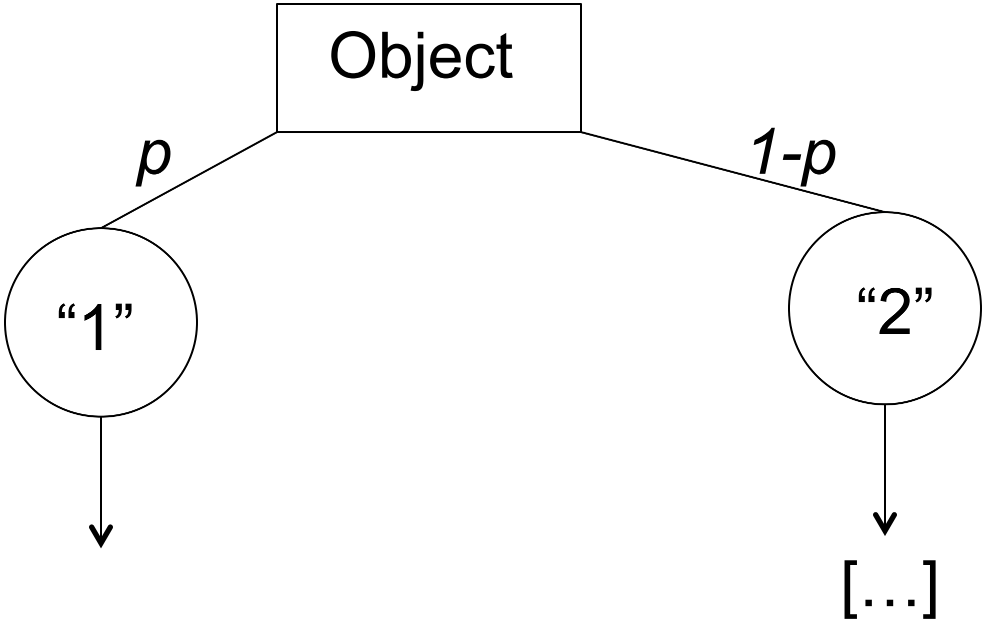
- A cognitive model of the preceding joint distribution of rater responses might include parameters indexing the **true category frequencies, rater accuracy, and category difficulty**. The joint distrn has 3 df (4 probs that sum to 1), so a model can have **at most 3** estimable parameters.
- See *'ho-1-narratives.pdf'*, Sec. 7, for possible models. We consider **Model 7.1**, which allows us to estimate p = P("1"), as well as the 'sensitivity' of each rater (for a total of 3 parameters).
- κ is approx 'accuracy'; but P(R1 = '1') and P(R2 = '1') depend on p *and* the guessing rate.

Rater 1, Category “1”



P(Resp = 1 | “1”) = $w_1 + .5(1 – w_1) = .5(1 + w_1)$,
P(Resp = 2 | “1”) = $.5(1 – w_1)$.
Similarly for Rater 2 or Category “2”.

Responses of Rater 1 and Rater 2 are independent, *given the Category* (“1” or “2”) – the *conditional independence* model. On a random trial:



P(R1 & R2 choose 1 | “1”) =
P(R1 chooses 1 | “1”)*P(R2 chooses 1 | “1”)
= [$.5(1 + w_1)$]*[$.5(1 + w_2)$].

P(R1 & R2 choose 1 | “2”) =
P(R1 chooses 1 | “2”)*P(R2 chooses 1 | “2”)
= $[.5(1 - w_1)]*[.5(1 - w_2)]$.

Therefore, on a random trial,

P(R1 & R2 choose 1) = P(“1”)*P(R1 & R2 choose 1 | “1”)
+ P(“2”)*P(R1 & R2 choose 1 | “2”)
= $p*[.5(1 + w_1)]*[.5(1 + w_2)] + (1 – p)*[.5(1 - w_1)]*[.5(1 - w_2)]$
= $.25(1 - w_1)(1 - w_2) + .5p(w_1 + w_2)$
== p_{11}.

Similarly for other joint probs. These equations are used to get ML parameter estimates – see ‘skapreliab2.r’.

Outline of Lecture 1.3

- A problem with *κ* when a category is rare
- Cognitive Models offer a solution, but are also useful in their own right
- Maximum Likelihood (ML) estimation of model parameters; the bootstrap
- Bayesian estimation and MCMC
- Useful code in ‘skapreliab2.r’
- Steve S’s problem, when an ‘expert’ and a ‘novice’ are the 2 raters; see ‘skapreliab2s.r’

Model analysis

The data are the joint freqs, (n_{11}, n_{12}, n_{21}, n_{22}). The likelihood of the data is:

$$l = k\left(p_{11}^{n_{11}}\right)\left(p_{12}^{n_{12}}\right)\left(p_{21}^{n_{21}}\right)\left(p_{22}^{n_{22}}\right),$$

where k is a constant that is independent of the p_{ij}'s. `l` is calculated by the function, `loglik2()`. This function is then used by `param.boot1()` in the **minimization** of $-L = -\log(l)$, to get ML estimates of the model parameters. We use the optimizer, `nlminb()`.

Also, standard errors can be obtained by using the **bootstrap**.

‘skapreliab2.r’

```
loglik2 = function(th, dat) { # dat =
(n11,n12,n21,n22); compute -log(likelihood) for
minimisation in nlminb()

   p0 = th[1]; w1 = th[2]; w2 = th[3]
   p11 = .25*((1 - w1)*(1 - w2)) + .5*p0*(w1 + w2)
   p12 = .25*((1 - w1)*(1 + w2)) + .5*p0*(w1 - w2)
   p21 = .25*((1 - w2)*(1 + w1)) + .5*p0*(w2 - w1)
   p22 = .25*((1 - w1)*(1 - w2)) + .5*(1-p0)*(w1 + w2)

   -(dat[1]*log(p11) + dat[2]*log(p12) +
dat[3]*log(p21) + dat[4]*log(p22) )

}
```

‘skapreliab2.r’ and `nlminb()`

```
param.boot1 = function(data) {# data =
   (n11,n12,n21,n22); est kappa and ML theta

   rs0 = kappa0(data)  # Cohen’s kappa
   rs00 = nlminb(start=c(0.5,0.5,0.5),
   loglik2, dat = data, lower=c(0,0,0),
   upper=c(1,1,1))$par  # ML theta
   rs1 = c(rs00, rs0)
   names(rs1) = c("p","w1","w2","kappa")
   round(rs1, 4)
}
```

Used in the bootstrap.

Model-based defn of Reliability

From the above equations, the proportion of agreement, p_a, is given by:

$$p_a = p_{11} + p_{22} = 0.5(1 + w_1 w_2).$$

Thus p_a is independent of prior prob, p, and depends only on the (geometric) average of the raters' sensitivities. If we were to *define* reliability as this average sensitivity, then:

$$\text{Reliability}, \ \omega \equiv (w_1 w_2)^{1/2} = (2p_a - 1)^{1/2}.$$

For the data, (96, 2, 2, 0), we saw that $\kappa = -.02$. It can be seen, however, that $\omega = 0.92$!

Output from ‘skapreliab2.r’

```
Crosstab freqs:  96 2 2 0
Cohens kappa: -0.0204

 Parameter estimates

  p      w1       w2    kappa
 1.0000  0.9600  0.9600 -0.0204

Bootstrap stats:
         mean  median      se  q.025 q.975
p      1.0000  1.0000 0.0000  1.000     1
w1     0.9599  0.9600 0.0276  0.900     1
w2     0.9604  0.9600 0.0274  0.900     1
kappa -0.0155 -0.0152 0.0114 -0.039     0
```

Bayesian Analysis

Let ϑ denote the parameters, (p, w_1, w_2), D the data, and *f()* the probability density function. Then, by Bayes Theorem [P(A&B)=P(A|B)P(B) =P(B|A)P(A)],

$$f(\theta \mid D) \propto f(D \mid \theta) f(\theta),$$

where $f(\vartheta \mid D)$ is the unnormalised **posterior density**, $f(D \mid \vartheta)$ is the likelihood of the data (denoted earlier by l), and $f(\vartheta)$ is the prior distribution. When $f(\vartheta)$ is the Uniform distribution on (0, 1), $f(\vartheta)$ = 1, and

$$f(\theta \mid D) \propto f(D \mid \theta) = l.$$

Bayesian Analysis

$$f(\theta \mid D) \propto f(D \mid \theta) = l.$$

Thus the value of ϑ that maximizes *l*, i.e., the ML estimate, also maximizes the posterior density, i.e., is the **mode** of the posterior distribution. The posterior mode is a good Bayesian estimate of ϑ, as are the posterior mean and median. *So, under Uniform priors, ML estimates and Bayesian estimates are similar*.

Often it is convenient to use MCMC to get Bayesian estimates (see ‘skapreliab2.r’).

‘skapreliab2.r’ and metrop()

```
library(mcmc)
param.mcmc1 = function(data, R) { # Bayesian
   estimates using MCMC
   rs1 = metrop(loglikmc1, rep(0, 3), nbatch
   = R, dat = data)
   rs2 = apply(rs1$batch, 2, boot.summary1)
   rs2 = t(rs2)
   rs3 = round(exp(rs2)/(1 + exp(rs2)), 4)
   # Transform from logit to probability scale
   rownames(rs3) = c("p", "w1", "w2")
   colnames(rs3) = c("mean", "median", "se",
   "q.025", "q.975")
   rs3
}
```

Output from ‘skapreliab2.r’

Parameter estimates using Bayesian model and MCMC (similar to ML estimates!)

```
     mean median     se  q.025  q.975
p  1.0000 1.0000 0.9907 0.9508 1.0000
w1 0.9668 0.9647 0.7027 0.8732 0.9933
w2 0.9646 0.9631 0.7208 0.8083 0.9942
```

(The se’s seem too large; perhaps there’s a bug in the script.)

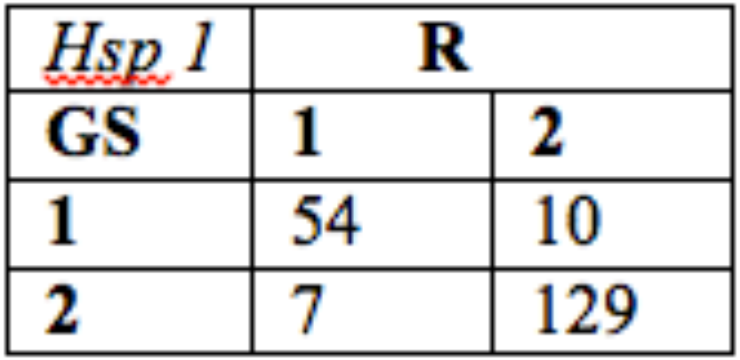
Outline of Lecture 1.3

- A problem with *κ* when a category is rare
- Cognitive Models offer a solution, but are also useful in their own right
- Maximum Likelihood (ML) estimation of model parameters; the bootstrap
- Bayesian estimation and MCMC
- Steve S's problem, when an 'expert' and a 'novice' are the 2 raters; see 'skapreliab2s.r'

Reliability for ‘expert-novice’ pairs

(Continuing to explore what we might do when an off-the-shelf statistic, *κ*, is not satisfactory)

- Steve S wrote a script for automated detection of any reference to ‘Patient is not taking the medication as prescribed’ in a Patient’s medical records. This is the Gold Standard (GS). Because GS is expensive, we need to train cheaper raters to mimic GS.
- Raters (R) are compared singly to GS, and there is an asymmetry between R and GS. ‘Inter-rater agreement’ is then misleading, and the interpretation of Cohen’s *κ* is unclear.
- Model 6.3 in *Handout 1* can be applied here, but let us use a threshold-type model (Model 6.4) that is similar to Model 6.1. **How to characterise the rater, R**?



Hsp 1	**R**	
GS	**1**	**2**
1	54	10
2	7	129

- GS’s response to an item is regarded as the ‘truth’; p = P(GS responds ‘1’) is the true rate of ‘1’s. Let w be the probability that each item (a “1” or a “2”) is above threshold, i.e., is clearly perceived, in which case R’s response is the true category. If an item is below threshold, the rater guesses and labels it “1” with probability g (which may differ from 0.5). Thus our 3 parameters are p, w and g.
- Assume that g has a **Beta(*a*, *a*)** prior, for some a, and that p and w have the **Uniform** prior on (0, 1).

- The Beta prior distrn for g is assumed to be:

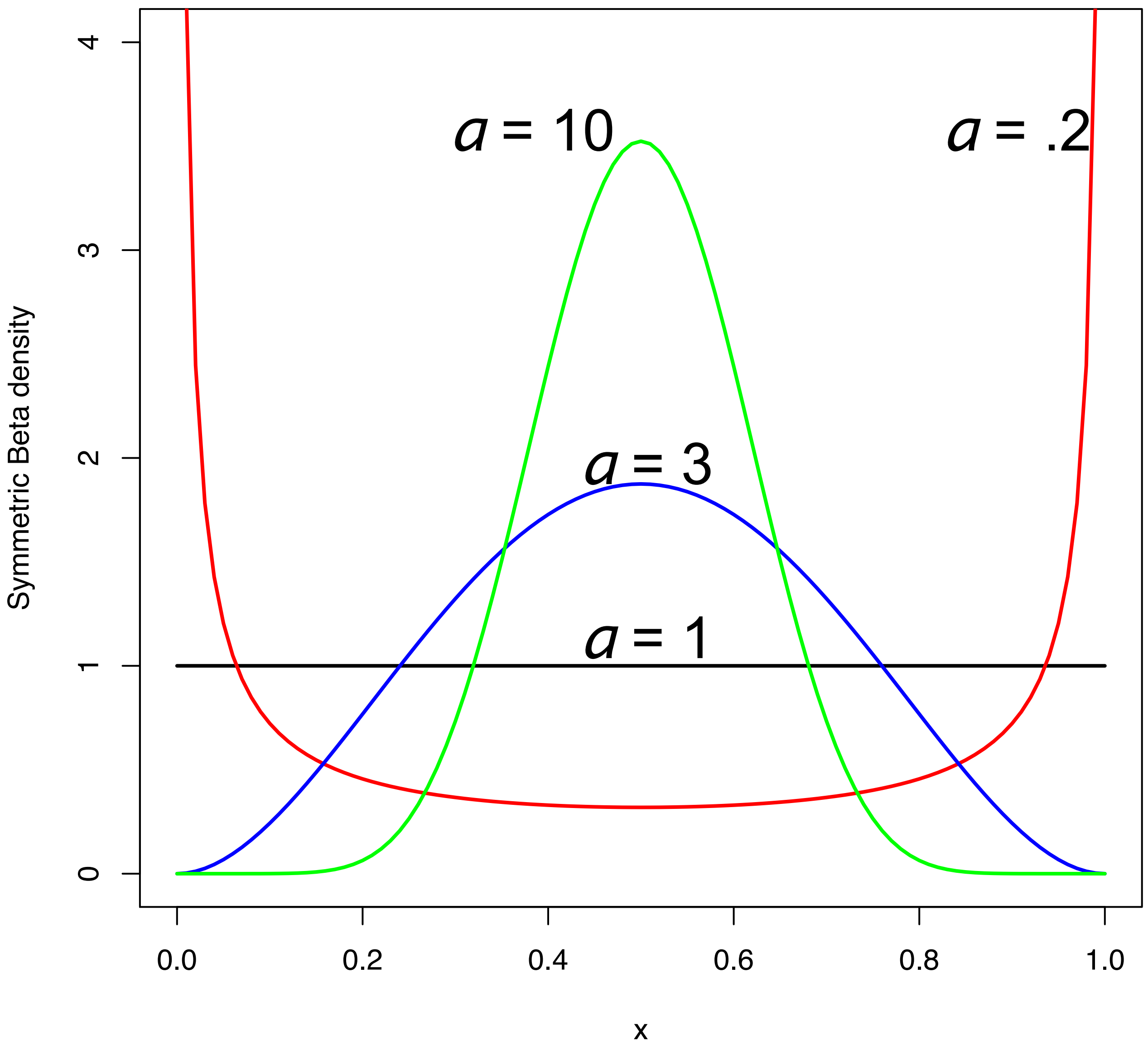
$$h(g; a, b) = k' g^{a-1}(1-g)^{b-1}.$$

- For simplicity, we assume *a = b*, which implies that the mean = 0.5, and the variance = 1/(8*a*+4).
- That is, our **prior** uncertainty about *g* is **centered on 0.5**. As *a* **increases**, this uncertainty becomes **more concentrated** about 0.5, and our posterior estimate of *g* will be drawn closer to 0.5. As before,

$$f(\theta \mid D) \propto f(D \mid \theta) f(\theta) \propto f(D \mid \theta) h(g; a, a),$$

- where *a* is given.

Density functions for the Symmetric Beta(a, a) distrn



‘skapreliab2s’

```
loglik.bayes3 = function(th, dat, a) {

  p = th[1]; w = th[2]; g = th[3]
  lprior.g = (a-1)*(log(g)+log(1-g))    # log
     prior for g; prior is Beta(a,a)
  p11 = p*(w + g*(1-w))
  p12 = p*(1-g)*(1-w)
  p21 = (1-p)*(1-w)*g
  p22 = (1-p)*(w + (1-w)*(1-g))
  -(dat[1]*log(p11) + dat[2]*log(p12) +
  dat[3]*log(p21) + dat[4]*log(p22) +
  lprior.g)
}
```

‘skapreliab2s’

```
param.boot1 = function(data, a0) { # data =
(n11,n12,n21,n22); est kappa and ML estimate
of theta
   rs0 = kappa0(data)
   rs00 = nlminb(start=c(0.5,0.5,0.5),
   loglik.bayes3, dat = data, a = a0,
       lower=c(0,0,0), upper=c(1,1,1))$par
   rs1 = c(rs00, rs0)
   names(rs1) = c("p","w","g","kappa")
   round(rs1, 4)
}
```

Output from ‘skapreliab2s’

- Joint freqs in sample 1: (54, 10, 7, 129)

a	***p***	***w***	***g***	***kappa***
0.1	0.32	0.79	0.23	0.80
1.0	0.32	0.79	0.25	0.80
10	0.32	0.81	0.36	0.80

- As **prior** uncertainty becomes more concentrated around g = 0.5 (i.e., as a increases), the prior mean (0.5) is more heavily weighted than the estimate based only on the data, and the posterior estimate of g approaches 0.5. The estimates of p and w are unaffected by a.

Model-based description of R

- The estimate of *p* is based only on GS's responses, not surprisingly because GS is the gold standard.
- Each rater (R) is characterised by a 'sensitivity', *w*, and a response bias, *g*, to say '1' when the category information is ambiguous.
- Training should increase *w*. What about *g*?! Shd the goal be *g* = 0.5?

Optimal *g*?

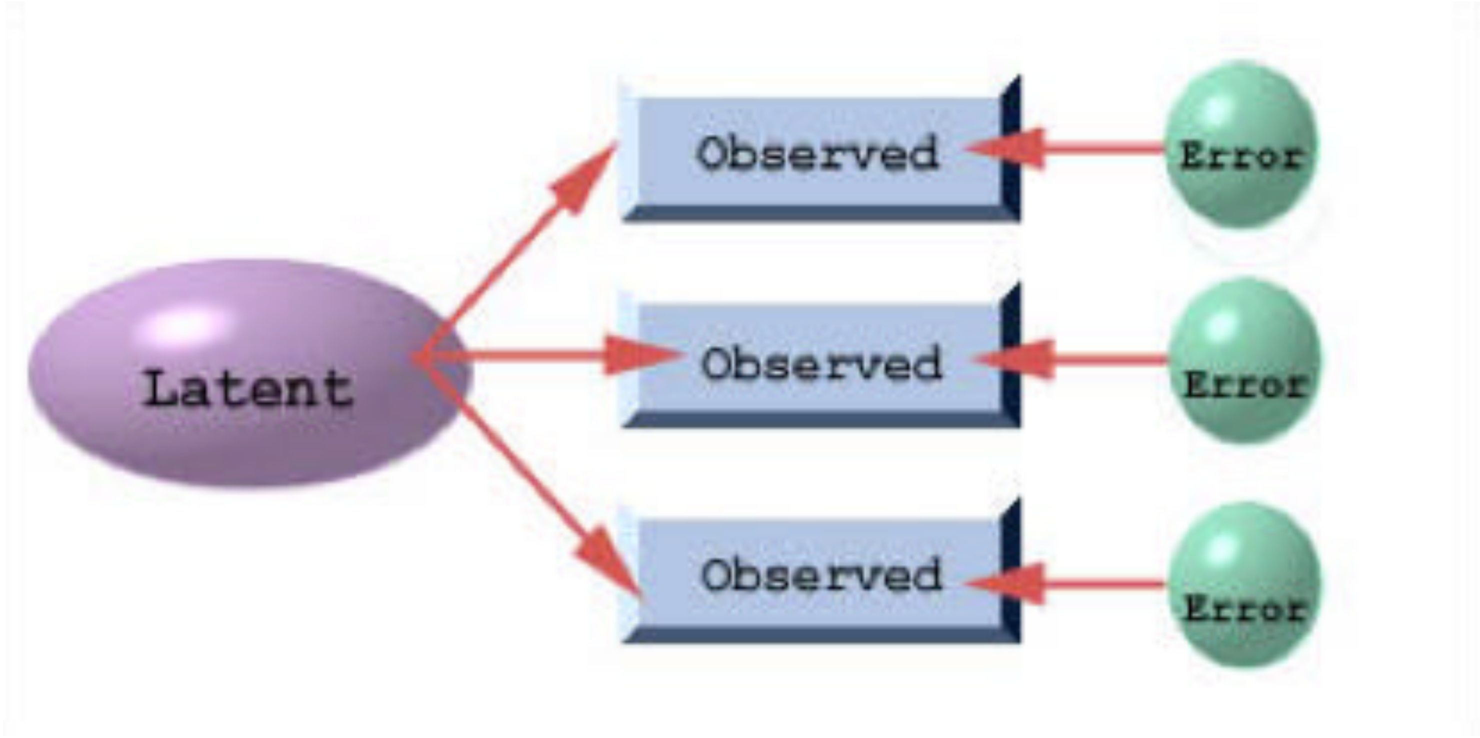
- Because GS is the gold standard, it is tempting to use p_a as an adequate measure of R's 'ability'. However (and as a corrective to this view), it can be shown that, according to Model 6.4,

$$p_a = [w + (1 - w)(1 - p)] + [(1 - w)(2p - 1)] * g.$$

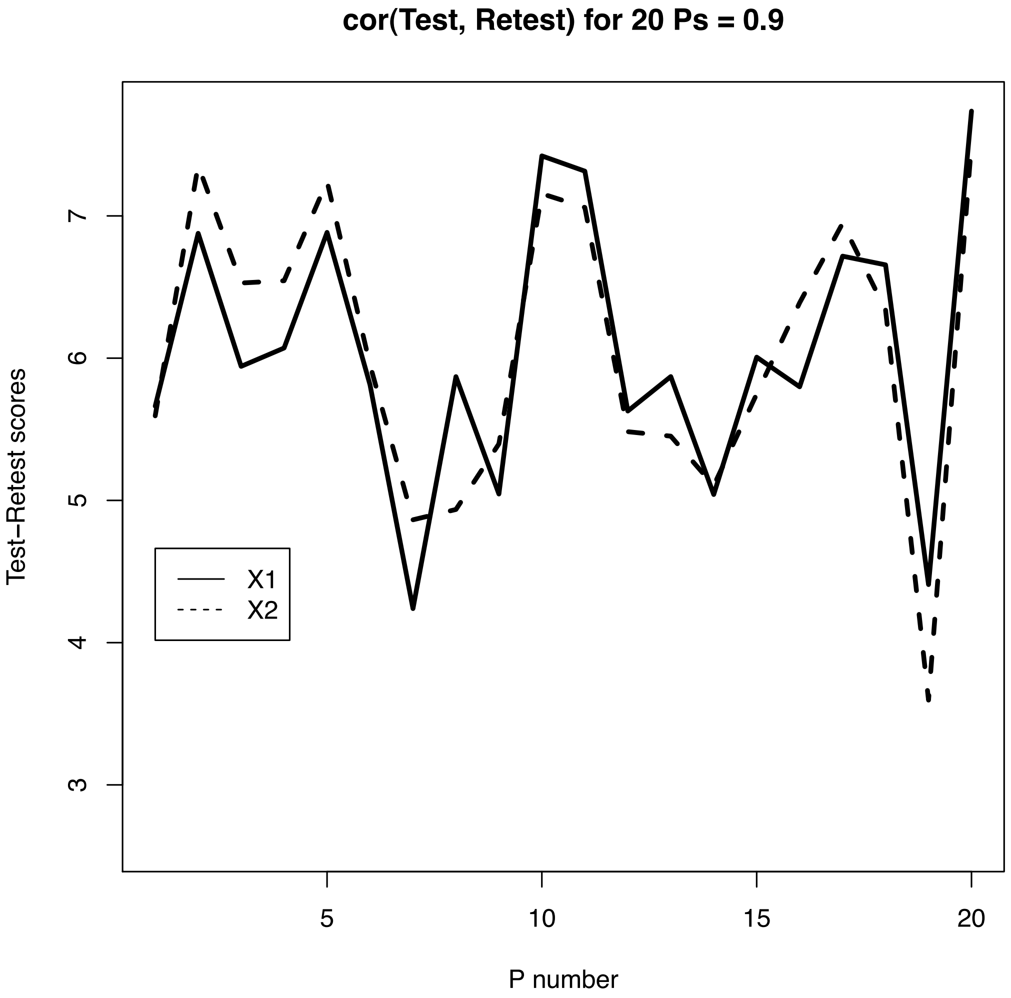
- First, p_a depends on all 3 params – it is not a pure index of 'sensitivity', *w*.
- Second, the effect of *g* on p_a depends on the **sign** of $(2p - 1)$: when category '1' is rare ($p < 0.5$, and $2p - 1 < 0$), p_a decreases as *g* increases. The opposite is true when category '2' is rare.

Pivot to theory-based approaches to reliability

- **Given data** in a certain ‘format’, we know how to compute reliability.
- Now consider two **theory-based** approaches within a very general model, **Test Theory**.
- First, assume *observed*, *X*, is the sum of a *latent* (or *true*) score, *T*, plus an error, *e*: $X = T + e, X' = T + e'$.
- ***Test-Retest*** *reliability = corr(X, X’)* across persons. I think of test-retest reliability as the **most concrete form** of the abstract concept.
- Consider 2 visual displays of this form



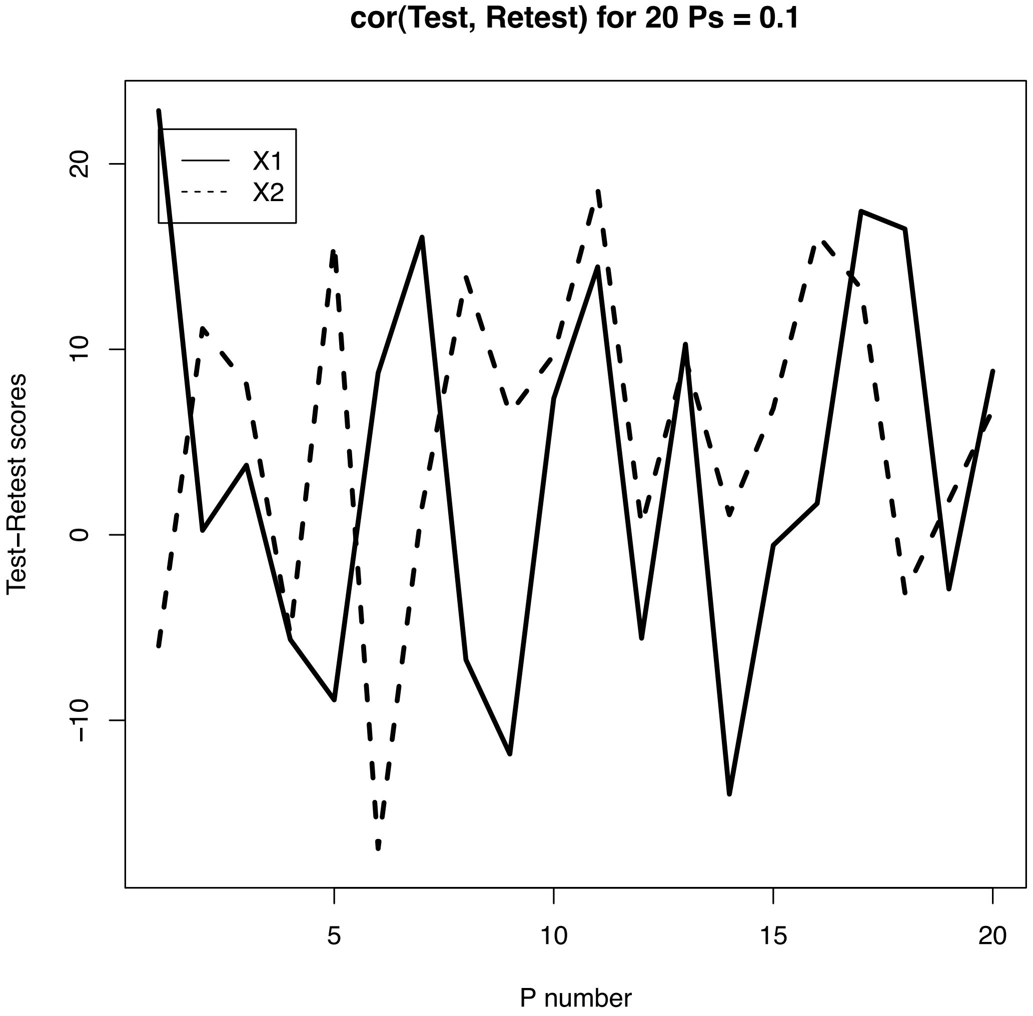
For a given person, *Latent* is fixed; successive measures of *Observed* (e.g., *X, X', X''*) on this person are different because of different *Error* terms. Across persons, *Latent* varies, and this variation induces a correlation, $r_{XX'}$, between *X* and *X'*, known as the *test-retest reliability*.



Think of ‘test’ score, X1 (or X) as Rater 1’s score, and X2 (or X’) as R2’s score; P is the ‘object’ being rated.

X = T + e
is a special case of
Y = a + bX + e,
i.e., simple regression.

So r^2 *(=* $r(X,T)^2$) is the proportion of *X*-variance that is explained by *T*. This links ‘correl’ to ‘reliability’.



In this 2nd case of low test-retest reliability, there is a **significant rater * object interaction**.

Internal Consistency

- In the second theory-based approach, a **'test'** consists of *n* **'items'**, and test score, *X*, is the sum of the item scores, *Y*:

  $$X = Y_1 + \ldots + Y_n$$

  (instead of the sum of true score plus error:

  $$X = T + e)$$

- Define *reliability* as ***internal consistency*** of a 'test', which is based on the correlations among Y_i and Y_j across participants.

	Items			Total, X
	1	2	3 …	
P1	8.0	6.2	8.4	22.6
P2	7.4	9.8	8.5	25.7
P3	10.4	7.5	9.1	…
P4	4.6	3.7	5.6	
P5	9.5	9.7	9.9	
P6	9.2	8.6	8.6	
P7	5.3	4.7	5.5	
P8	6.6	7.9	6.9	
P9	4.2	6.9	3.6	
P10	5.4	6.4	6.9	
P11	6.9	6.9	10.2	
P12	10.6	10.6	11.6	
P13	11.5	11.6	11.3	
P14	6.3	8.2	6.8	
P15	7.5	9.7	9.0	26.2

The test is reliable if cor(1, 2), cor(1,3), etc. are high; i.e., if the items “hang together”.

Summary of Theory-based results

- Test-retest *reliability = var(T)/var(X)*, **proportion** of **observed score** variance explained by **true score** variance.
- Reliability of a test increases with the *length*, *n*, and with the *internal consistency*, *ρ*, of the test (Spearman-Brown formula), where *ρ* is the correl between items across persons.
- [To see the generality of this model, note we can replace 'test' by 'method', 'item' by 'rater', and 'person' by 'object'.]

Summary (cont'd)

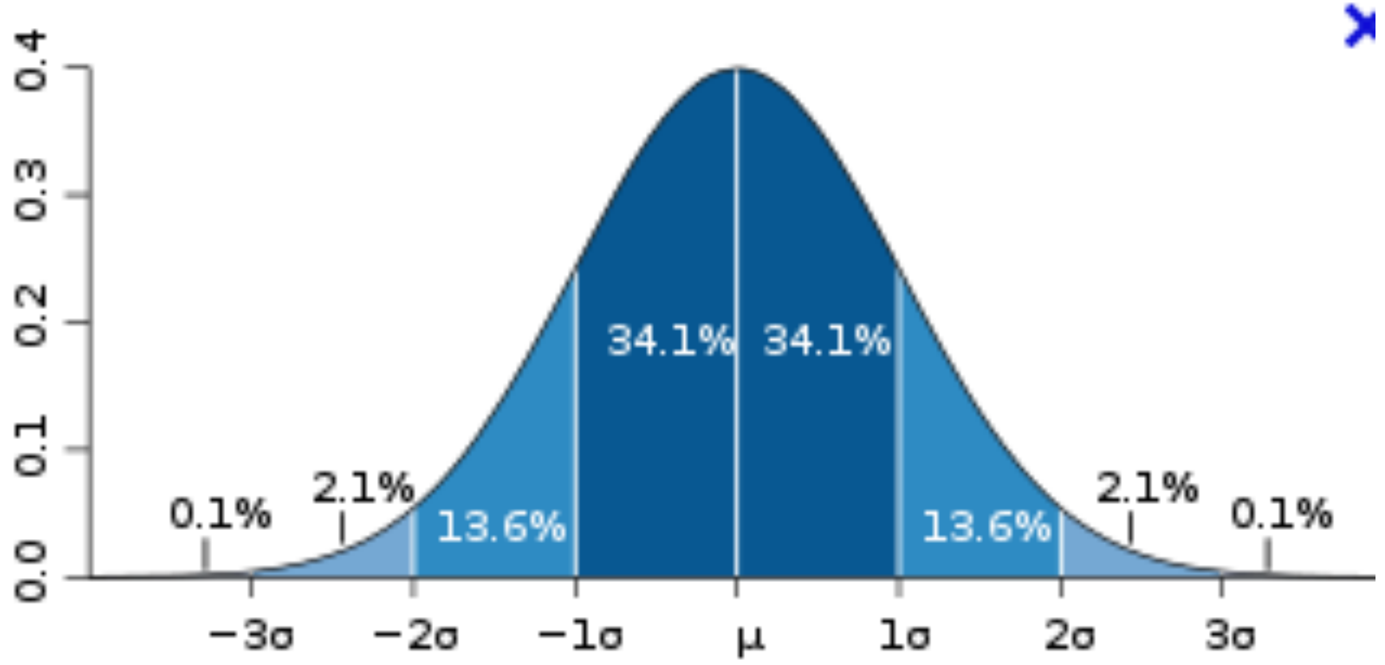
- Reliability is reduced when we reduce the range (or variance) of *X*
- The estimate, r_{XY}, of a hypothetical correlation, r_{AY}, between some 'interesting' variable, *Y*, and a *latent* construct, *A*, that is measured by *X*, increases as the reliability of *X* increases.
- But before the theories, consider the **purpose** of *causal models*, e.g., Test Theory and SEM, and *latent variables,* given that 'correlation doesn't imply causation'?
- We need causal models in theory and policy settings. Often we can't do experiments and must be content with correlations.

The case for SEM by *Wiki*

- http://en.wikipedia.org/wiki/Structural_equation_modeling
- SEM: A statistical technique for testing and estimating *causal relations* using a combination of statistical data and qualitative causal assumptions
- Strengths include the inclusion of *latent variables*. These are variables which are not measured directly, but are estimated in the model from several measured variables, each of which 'taps into' the latent variables.
- This allows separate estimation of *unreliability of measurement* (noise) in the model, and the *structural relations* among latent variables.

- Caution: In *Applied Multivariate Data Analysis* (Chap 13), by BS Everitt and G Dunn (2001)
- Causal models "are best seen as convenient mathematical fictions which describe the investigator's belief about the causal structure of a set of variables. … Essentially, so-called causal models simply provide a parsimonious description of a set of correlations."

- “Latent variables are … hypothetical constructs invented by a scientist for the purpose of understanding some research area of interest, and for which there exists no operational method for direct measurement. … They serve to synthesize and summarize the properties of observed variables.
- Latent variables are as real as their predictive consequences are valid. … The justification for postulating latent variables is their **theoretical utility** rather than their reality.”



Model: "***X*** has mean, *μ,* and s.d.*, σ"* is equivalent to

"X = μ + ε, where ***ε*** has mean*, 0,* and s.d.*, σ".*

We can call *μ* the *true score, T,* of *X*, and ε the *measurement error* in/of *X*. More typically,

$X = T + \varepsilon,$

where *T* varies across persons with some variance, σ_t^2.

- An easy case*:* *X* is a Rater's estimate of the **length** of an object. We have **objective** measures on this scale (ins, ft.), and can say: **High rater reliability = Low σ^2.** Can also define Rater **bias** as the difference between mean of *X* and objective length.
  - the meaning of 'low variance' depends on context.
- A hard case: *X* is a person's score on a personality test. Is the test reliable? We have no **objective** measures for this scale, and 'reliability' is more problematic.

- *X* is modeled as the sum of the person' s true score, μ (typically written at t_i for *i*' th person), plus an **independent** error of measurement, ε: $X_i = t_i + \varepsilon_i$.
- High reliability = Low σ_e^2, **relative** to the variance of the true scores across persons.
- $Var(t_i) = \sigma_t^2$; $var(\varepsilon_i) = \sigma_e^2$;
- *Reliability,*

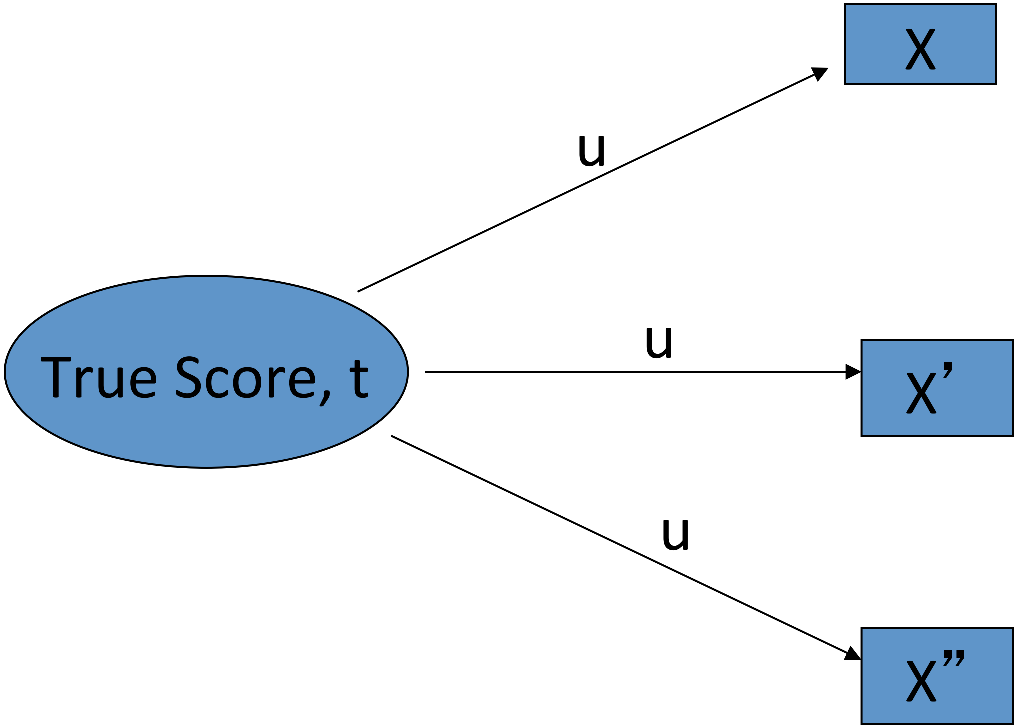
$$r = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_e^2}.$$

$$r = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_e^2}.$$

- We can't observe the true score or the error; so we cannot directly calculate the two variances. We need more data.
- Let *X* and *X'* be **two measures from the same person** on the test. $r_{XX'}$ is the correl across persons, and can be calculated. Relating $r_{XX'}$ to 'reliability' is the key insight of Test Theory.

Corr(X, X') = $u^2 \equiv r_{XX'}$ = Test-Retest reliability.

$u \equiv r_{Xt}$ ‘large’ implies ‘high’ reliability.



Theorem: If $\boldsymbol{t \rightarrow X}$ ***and*** $\boldsymbol{t \rightarrow X'}$, then $\boldsymbol{r_{XX'} = r_{tX} * r_{tX'} = u^2}$

- **Proof**: Assume that *X, X', t* and *e* are **standardized** variables.
- The regression of *X* on *t*, is $X = ut + (1\text{-}u^2)^{1/2}e$.
- The regression of *X'* on *t*, is $X' = ut + (1\text{-}u^2)^{1/2}e'$.
- Multiplying these 2 eqns, we get:

  $XX' = u^2t^2 + ut*(1\text{-}u^2)^{1/2}e + \ldots$
- On taking expected values,

  $E(XX') = u^2E(t^2) + E[ut*(1\text{-}u^2)^{1/2}e + \ldots] = u^2,$

  because $E(t^2) = 1$, and the variables inside the *[]* brackets all have expected value = 0.

Completion of Proof

- Because X and X' are standardized,

  $r_{XX'} = E(XX') = u^2$, as stated earlier.

- This result will be used many times, e.g., in deriving correlations among observable variables in **Factor Analysis**.
- Note that $u = corr(t, X)$. Thus, in the 'regression',

  $X = bt + e'$,

- u^2 (i.e., "R^2") is the proportion of variance in X that is explained by t. Thus, **reliability, $r_{XX'}$, is the ratio of true score variance to total observed variance in *X***.
- Another proof follows.

- $X = t + \varepsilon$; $X' = t + \varepsilon'$.
- True score is the same, but error of meas differs across testing occasions. To simplify, assume
- $E(t) = 0$. This implies $E(X) = 0$, var(t) = E(t^2), and var(X) = E(X^2).
- var(X) = var(X') = var(t) + var(ε);
- $XX' = t^2 + t\varepsilon + t\varepsilon' + \varepsilon\varepsilon'$. So
- Cov(X, X') = E(XX') = E(t^2) = var(t) = σ_t^2

$$r_{XX'} = \frac{\operatorname{cov}(X, X')}{\sqrt{\operatorname{var}(X)\operatorname{var}(X')}} = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_e^2}$$

$$r_{XX'} = \frac{\operatorname{cov}(X, X')}{\sqrt{\operatorname{var}(X)\operatorname{var}(X')}} = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_e^2}$$

- This shows that, even if we can’t observe true scores and errors, we can estimate the reliability if we have test and re-test data.
- The various definitions of ‘reliability’ can all be related to this fundamental definition.
- Next we turn to principles for test construction.