Lecture notes to Stock and Watson chapter 3

Review of elementary statistics

Tore Schweder

August 2009

Statistical inference

Statistical inference the art/science of drawing **conclusions** from observed **data** in view of **theory** by way of a **model**.

- The model, as determined by the value of the parameter of the model, describes the **data generating process.** It gives the probability distribution of the data *ex ante* for given state of the system/reality. In parametric models (which SW deals with) the state of the system is determined by a (finite dimensional) **parameter.**
- The model is chosen according to the purpose of the study, e.g. the **questions** - in view of **substantive (economic) theory**: the mechanism behind the data; and **statistical theory**: guiding towards effective inference.

Model

The statistical model bridges the gap between economic theory and data - allowing the relevant information to be extracted from the data.

- The model must reflect the basic and important pattern of the data.
- Example of wrong model leading to grat damage: The risk in financial assets have long-tailed distributions, and are positively correlated. The Gaussian copula is a model for the joint risk distribution of two assets X, Y with marginal cdfs $F(x) = P(X \leq x), G(y) = P(Y \leq y)$

$$P(X \leq x, Y \leq y) = P\left(Z_1 \leq \Phi^{-1}(F(x)), Z_2 \leq \Phi^{-1}(G(y))\right)$$

(Z_1, Z_2) is satndard bivariate normally distributed with correlation ρ, Φ is the standard normal cdf. This model was widely used for estimating the correlation, and for setting prices of multiple securities. It purely reflects the real association between financial risks, and is partly blamed for the global financial crisis - as a "receipe for disaster".

Types of statistical methods (questions, answers)

Given a model **M** with a parameter θ and data D.

- **Point estimation** (what is the value of θ? a numeric value for θ!). Method for informed guessing (based on D and **M**) of the numerical value of θ.
- **Confidence interval estimation (**what is the value of θ? an interval for θ associated with a **level of confidence** $1-\alpha$!**).** Methods for informed guessing of an interval that in repeated studies covers the true value of θ with controlled probability $1-\alpha$**.**
- **Hypothesis testing** (how much confidence do we have in a statement about θ? a number measuring the evidence!). H_0 and H_1 covers all values of θ. The $\mathbf{p-value}$ measures the evidence for H_0. It is the probability in a future experiment of obtaining data at least as contrary to the null hypothesis as those observed. Reject H_0 at confidence level α if $\mathbf{p-value} < \alpha$. *Ex ante* $\mathbf{p-value}$ is a random variable $\overset{ST}{\geq} U(0,1)$ under H_0 and $\overset{ST}{<} U(0,1)$ under H_1.

Example

Questions about desired number of children X for potential beginning master students in economics in Oslo around 2008:

1. What is $EX = \mu$?
2. Is $p = P(X = 0) > 1/2$?

- **D** : numbers of children for students in the auditorium, $X_1, \cdots, X_n$.
  M : $X_1, \cdots, X_n$ are iid copies of X. Random sample from an infinite population!

1. Point estimation: $\widehat{\mu} = \overline{X}$. $SE(\widehat{\mu}) = \widehat{\sigma}/\sqrt{n} = s/\sqrt{n}$. 95% confidence interval: $\widehat{\mu} \pm 1.64 \cdot SE(\widehat{\mu})$
2. $H_0 : p \leq 1/2 \quad H_1 : p > 1/2$. Test statistic: $\widehat{p} = \frac{\#X_i=0}{n}$. Null distribution: $n\widehat{p}$ binomially distributed $(n, p = 1/2)$
   $\Rightarrow T = \frac{\widehat{p}-0.5}{\sqrt{0.5 \cdot 0.5/n}}$ is approximately $N(0,1)$.
   $\mathbf{p - value} = 1 - F_{N(0,1)}(T)$.

Constructed population - simulation

D :

x (number of children)	0	1	2	3	4	sum
$\#\{X_i = x\}$	39	20	3	1	1	64

1. $\widehat{\mu} = 0.516$, $SE(\widehat{\mu}) = 0.797/\sqrt{64}$, 90% confidence interval: $(0.352, \quad 0.679)$.
2. $\widehat{p} = 39/64 = 0.609$, $T = \frac{\widehat{p}-0.5}{\sqrt{0.5 \cdot 0.5/n}} = \frac{0.609-0.5}{1/16} = 1.75$, $\quad \mathbf{p-value} = 1 - F_{N(0,1)}(1.75) = 0.04$

- Population:
  $\mu = 0.57$, $\quad \sigma = sd(X) = 0.875$, $\quad p = P(X = 0) = 0.6$. $n = 64$.
-

x	0	1	2	3	4	sum
$f(x) = P(X = x)$	.6	.3	.05	.03	.02	1
$64 \cdot f(x)$	38.4	19.2	3.20	1.92	1.28	64

Properties of the confidence interval method (constructed example)

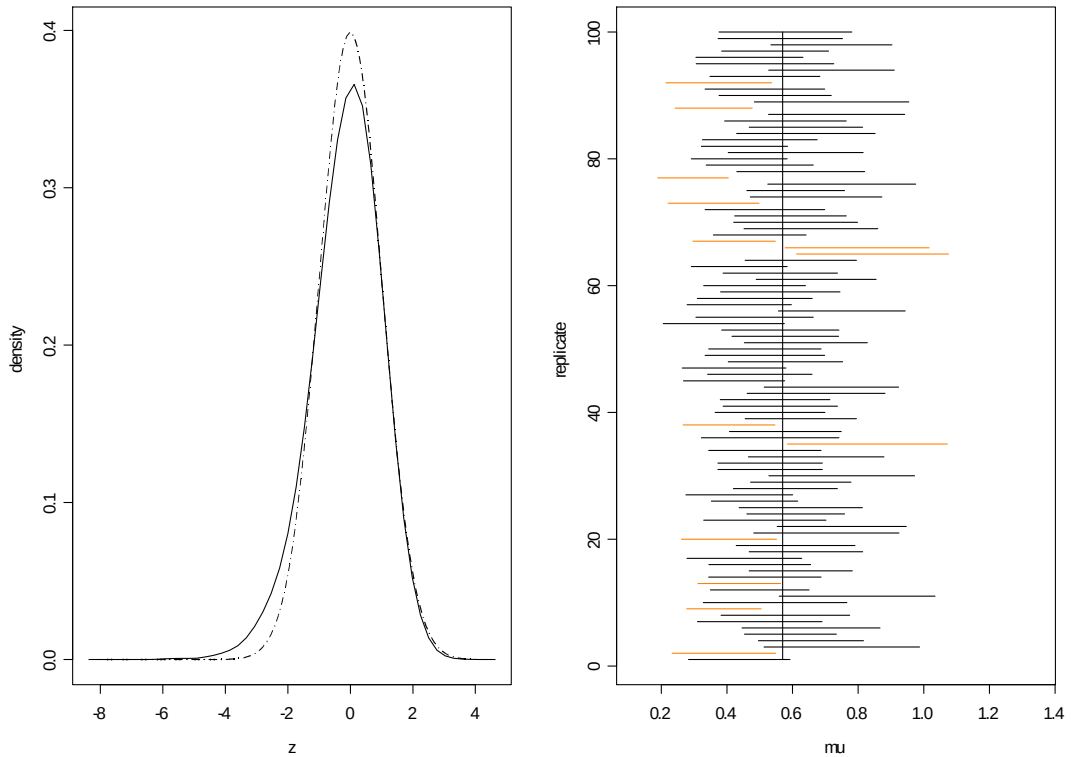

Figure: Left: density of $(\widehat{\mu} - \mu) / SE(\widehat{\mu})$ and of the standard normal distribution (broken line). Right: The 90% confidence intervals for 100 replicate simulations. Of 10,000 replicates, μ was covered in 8799 cases.

Properties of the test method (constructed example)

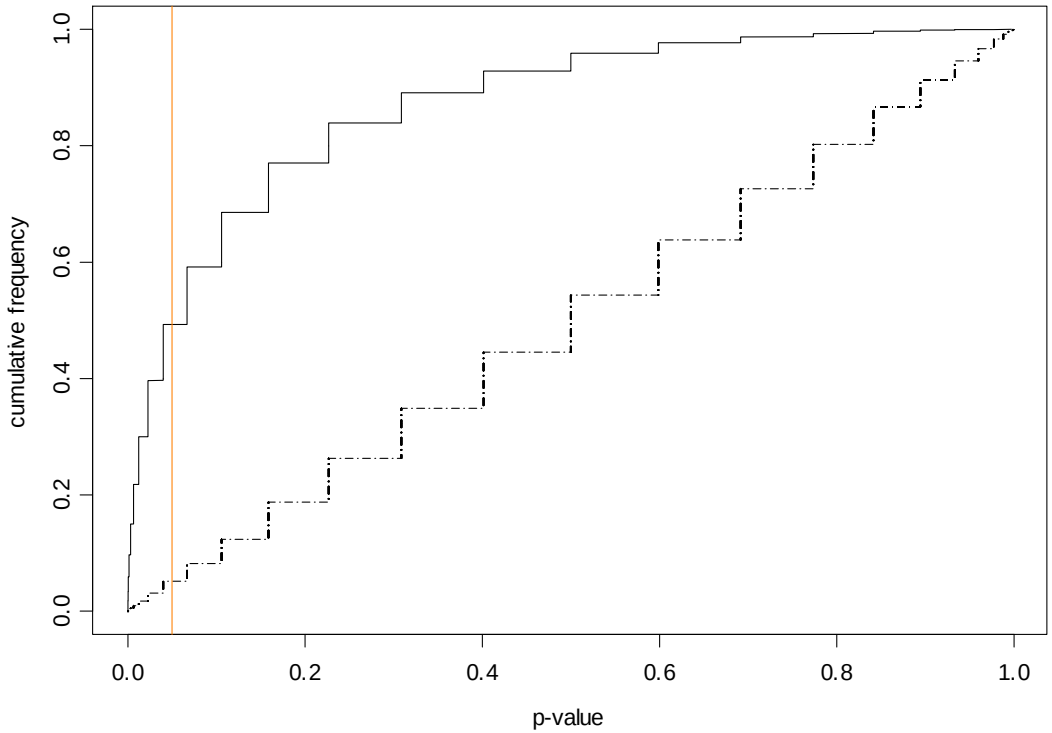

Figure: Cumulative distribution function of the p-value when $p = .6$ and in the border H_0 case $p = .5$ (broken line). The confidence level $\alpha = 0.05$ is indicated by the red vertical. The power at that level is 0.5.

Properties when n=43

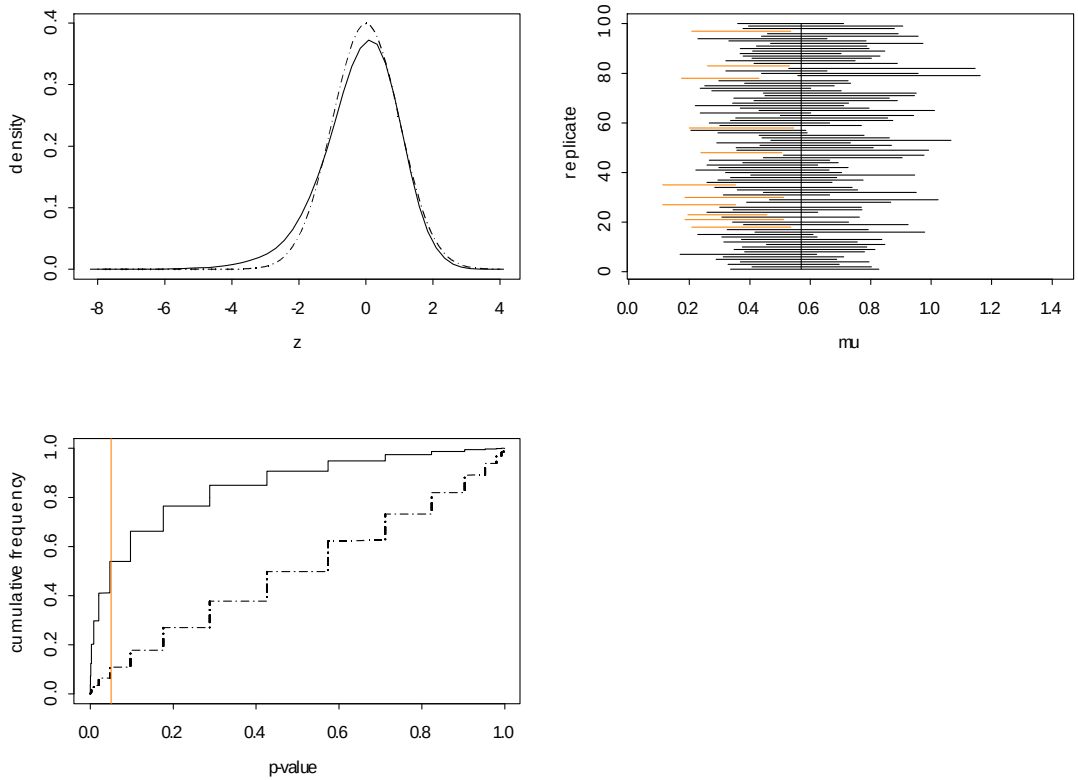

Figure: Same as Figures 1 and 2, but now with $n = 43$. Of the 10,000 simulated 90% confidence intervals 8755 covered μ.Less accurate confidence interval and less test power!

Problems to be done in class

- SW: 3.3
- SW: 3.13 with the addition: **c**. Let $\delta = \mu_{<20} - \mu_{\geq 20}$. What is the meaning of δ? Estimate δ and find $SE\left(\widehat{\delta}\right)$. Find a 99% confidence interval for δ.