

## **Interpreting John W. Tukey: WHERE MIGHT WE BE GOING?**

From the introduction:

*[There must be] a recognition of some analyses as approximations to -- or short-cut substitutes for -- more detailed, more complex, or more obviously appropriate analyses. How far we go down such paths does not have a single "correct" answer. Skill, experience, judgment have to be included in choosing an analysis whose conclusions can be widely accepted. What we do routinely in one decade may not be acceptable in a later decade.*

What is he trying to say?

I. Classical modeling paradigms

II. Basic ideas

A. Distributions

B. Classical point/interval estimation

C. Classical hypothesis testing

D. "New" approach

III. Illustration: CI for location parameter

## I. Classical Statistical Modeling paradigm

Have "model" for how physical phenomena behave

May be "simple" -- e.g.

*ozone concentration = f(wind speed, hour of day, season)*

or may be complicated with many differential equations

Most models involve *parameters* that we need to

-- estimate (point estimates)

-- assess their significance

-- understand in terms of uncertainties attached to them

Collect data related to our phenomenon and estimate parameters:

$$\textit{observed} = \textit{model} + \textit{error}$$

because model never fits *exactly*

Classical questions:

1. How can we use the observed data to estimate model parameters?
2. How good are these estimates, in light of the fact that different data will lead to different estimates?
3. What are the sources of uncertainty that would give rise to the different estimates?

## II. Statistical concepts

### A. Distributions

$$data = fit (model) + error (residual)$$

Ex: observe

$$y_i = f(x_i, t_i, p_i, \dots) + \varepsilon_i$$

Same  $x_i, t_i, p_i \Rightarrow$  same  $f_i$ , but  $\varepsilon_i$  varies ("random")

If we cannot nail down  $\varepsilon_i$  precisely, can we at least say something about how it is distributed -- typical values of  $\varepsilon_i$ , where they fall, say, 95% of the time, how often it stays within the limits  $\pm 0.10$ ?

N.B. Typically we assume that  $\text{ave}\{\varepsilon_i\}=0$ , because if it isn't, then we can just add in a constant to  $f_i$ .

Commonly used model for how  $\varepsilon_i$  is distributed:

$\varepsilon_i \sim N(0, \sigma^2) \equiv$  Gaussian (normal, bell-shaped curve) having mean 0 and standard deviation  $\sigma$  (95% of the  $\varepsilon_i$  like between  $\pm 1.96\sigma$  of 0).

Why Gaussian?

Usually because it leads to mathematically tractable solutions

Example: simplest model of all:

$$y_i = \mu + \varepsilon_i, \quad i=1, \dots, n. \quad (*)$$

Given observations  $y_1, \dots, y_n$ , what is our best estimate of  $\mu$ ? Sample mean  $\bar{y}$ ? Sample median  $y_M$ ?

If  $y_1, \dots, y_n \sim N(0, \sigma^2)$ , then  $\bar{y}$  is "better", in that:

$$\begin{aligned}\text{ave}(\bar{y}) &= \text{ave}(y_M) = \mu, \\ \text{var}(\bar{y}) &< \text{var}(y_M)\end{aligned}$$

How to estimate of  $\sigma^2 = \text{Var}(\varepsilon_i)$ ?

$$s^2 = \left[ \sum_{i=1}^n (y_i - \bar{y})^2 / (n-1) \right]$$

$$\text{avedev} = \left[ \sum_{i=1}^n |y_i - y_M| / n \right]$$

$$1.5\text{MAD} = 1.5 \times \text{median}\{|y_i - y_M|\}$$

*IF*  $\varepsilon_i \sim N(0, \sigma^2)$ , *AND* model (\*) holds, *THEN*  $s^2$  is better.

*BUT* if  $\varepsilon_i \sim (1-p)N(0, 1) + pN(0, 9)$ , *THEN* *avedev* is better than  $s^2$ ! (Tukey 1960)

JWT believes that there is an unrealistic reliance on the Gaussian distribution for our models to justify most of our procedures, and calls such situations "utopian."

## B. Shapes

"Shape" = distributional form, apart from constants of location and scale

- Gaussian shape  $N(\mu, \sigma^2)$ : all the same shape, differ only in location (mean) and scale (standard deviation)
- Exponential shape  $\text{Exp}(\mu, \lambda) = \lambda e^{-\lambda(x-\mu)}, x \geq \mu$
- "h-family":  $f_Y(\cdot)$ , where  $Y = Ze^{-hZ^2}$ ,  $Z \sim N(0,1)$

$h=0 \rightarrow Y = \text{standard Gaussian}$

$h < 0 \rightarrow \text{tails of } Y \text{ more } \textit{squeezed} \text{ than those of } Z$

$h > 0 \rightarrow \text{tails of } Y \text{ more } \textit{stretched} \text{ than those of } Z$

$$F_Y(y) = P\{Y \leq y\} = 0.95 \Rightarrow y = 1.96 e^{-h(1.96)^2}$$

-- so easy to compute percentiles for  $Y$

### C. Classical hypothesis testing

Example: Does caffeine raise one's blood pressure?

Procedure: Measure one's blood pressure (baseline), serve coffee with quantifiable amount of caffeine; remeasure blood pressure. Compare "BP after caffeine" with "BP before caffeine": is the difference,  $\delta$ , "roughly" zero, within measurement and sampling error?

$$H_0: \delta=0 \quad \text{vs} \quad H_1: \delta \neq 0$$

Assume  $H_0$  is true unless demonstrated otherwise

	True state of nature	
	$\delta=0$	$\delta \neq 0$
Decide $H_0$	correct	$\beta = \text{Type II error}$
Decide $H_1$	$\alpha = \text{Type I error}$	correct

Calculate average difference  $\bar{D}$  and the sample standard deviation  $s_D$ ; if  $|\bar{D}| > t_{n-1}(\alpha/2) \cdot s_D n^{-1/2}$ , then conclude that data are so extreme as to be unlikely to have occurred under  $H_0$ : choose  $H_1$ .

JWT philosophy: No one ever believes  $\delta=0$  exactly anyway. More likely, we seek to determine whether the data are convincing enough for us to conclude whether  $\delta > 0$  or  $\delta < 0$ , and by how much ("direction"). So there are really only 2 "decisions":

- (a) *Direction*:  $\delta > 0$  or  $\delta < 0$  can be ascertained, with some degree of confidence, from the data (there is always a chance that we are wrong);
- (b) Cannot confirm direction: Don't know if  $\delta > 0$  or  $\delta < 0$ .

**D. Classical estimation paradigm: how to estimate  $\delta$ :**

(1) Choose distribution=shape; e.g., Gaussian  $N(\mu, \sigma^2)$ .

(2) List variety of procedures

for  $\delta$ : sample mean =  $\bar{D}$ ; sample median  $D_M$

for  $\sigma$ :  $s_D = [\sum_{i=1}^n (D_i - \bar{D})^2 / (n-1)]^{1/2}$ ;

$1.5 \times \text{MAD} = 1.5 \cdot \text{med} |D_i - D_M|, \dots$

(3) Compare performance in terms of criteria:

bias: ideally,  $\text{ave}(\text{estimate}) = \text{target parameter}$

variance: e.g.,  $\text{Var}(\bar{D}) < \text{Var}(D_M)$

(4) Select the "best" -- e.g., lowest Mean squared error =  $(\text{bias}^2 + \text{variance})$

(5) Find a confidence interval for the parameter based on estimator; e.g.,  $\bar{D} \pm 2s_D n^{-1/2}$  for 95% confidence

JWT paradigm based on two facts:

- Data are observed, so configuration is known  $\{(D_i - \text{location})/\text{scale}\}$
- Shape is almost never known

JWT paradigm:

(1) Choose many shapes.

(2) Identify parameter of interest and confidence level.

(3) Calculate a confidence interval for the parameter for each selected shape, using the observed configuration.

(4) Take the convex cover of all such intervals (shortest interval that covers all).

Result: More emphasis on the uncertainty in the *shape*, which translates into uncertainty about the *parameter*

N.B. Still concerned with only *random* uncertainty (not systematic, or lack of fit), because that is the only kind of uncertainty that we know how to treat statistically.

A method for calculating confidence intervals:

Given  $n$  observations  $y_1, \dots, y_n$  from a density  $f$  having "center"  $\mu$  and "scale"  $\sigma$ , the likelihood is the product of the individual densities, as a function of the parameters:

$$\prod_{i=1}^n \sigma^{-1} \cdot f((y_i - \mu)/\sigma).$$

We do not know  $\mu$  or  $\sigma$ , so we replace them by  $t$  and  $s$ . Then  $d\mu = \sigma dt$  and  $d\sigma = ds$ . So, to get from  $dt ds$  to  $d\mu d\sigma$ , need to multiply by  $\sigma$ :  $d\mu = \sigma dt$ . So "density", as a function now not of  $\{y_i\}$  but of  $t, s$ ,

$$s^{-(n-1)} \prod_{i=1}^n f(s^{-1}(y_i - t)) dt ds$$

Now, integrate over  $s$  to get a "density" for  $t$ :

$$c_f(t) \propto \int_0^{\infty} s^{-(n-1)} \prod_{i=1}^n f(s^{-1}(y_i - t)) ds$$

For a valid density,  $\int_0^{\infty} c_f(t) dt = 1$ , need to divide by

$$A_f = \int_{-\infty}^{\infty} \left[ \int_0^{\infty} s^{-(n-1)} \prod_{i=1}^n f(s^{-1}(y_i - t)) ds \right] dt.$$

If we know  $f$ , then a "95% confidence interval" for  $t$  is

$$[ c_f^{-1}(0.025), c_f^{-1}(0.975) ] \quad (*)$$

(e.g. Student's  $t$  intervals, when  $f = \text{Gaussian}$ ).

But if we do *not* know  $f$ , then we could calculate the CI for various choices of  $f$ . What choices for  $f$ ?

(a) Various stretched-tailed densities ( $h > 0$ ); see Hoaglin 1985, p.479, 481-482. Calculate, by numerical integration,  $[c_f^{-1}(\alpha/2), c_f^{-1}(1-\alpha/2)]$ , for various  $h$ .

(b) Various data configurations  $\{(y_i - A) / B\}$ .

Ex:  $n=5$ , most "extreme" configurations:

$\{-1, -1, 0, 1, 1\}$

$\{-1, -1, -1, -1, 1\}$

$\{-1, 0, 0, 0, 1\}$ , etc.

Might consider other limits for other configurations;

e.g.,  $\{-1, -1/2, 0, 1/2, 1\}$

In general,  $-1 \leq a \leq b \leq c \leq 1$ , where  $b \leq 0$ . If we want to consider  $k$  possible values from which to choose  $a, b, c$  ( $k$  odd), then number of configurations satisfying the constraints is  $m(m+1)(4m-1)/6$ ,  $m=(k+1)/2$ .

$k=5$ ,  $\{a, b, c\} \in \{-1, -1/2, 0, 1/2, 1\}$

=> 22 configurations

$k=7$ ,  $\{a, b, c\} \in \{-1, -2/3, -1/3, 0, 1/3, 2/3, 1\}$

=> 50 configurations

$k=9$ ,  $\{a, b, c\} \in \{-1, -3/4, -1/2, -1/4, 0, 1/4, 1/2, 3/4, 1\}$

=> 95 configurations

Given a configuration, e.g.,  $\{-1, -1, 0, 0, 1\}$ , we can calculate  $A_f$ , a kind of "likelihood," where  $f = \phi \equiv$  Gaussian ( $h=0$ ) or a variety of other  $h$ -values. Suppose  $A_\phi \leq A_f$ .

- How much larger would  $A_f$  have to be before we reject  $f=\phi$  in favor of  $f =$  another choice, more consistent with the data?
- What values of  $h$  are plausible and consistent with the data (confidence interval)?

- For what value of  $h$ , say  $h_{\max}$ , is  $A_f$  maximum? Once we know that maximum value of  $h$ , we might ask whether it is so much larger than  $h=0$  (Gaussian) so as to reject the plausibility of the Gaussian model. Ex of a criterion:

$$2 \ln (A_h/A_\phi) = 5.9915 = X_2^2(0.95),$$

i.e., ratio  $\approx \exp(3) = 20.09$ . (why two degrees of freedom? 1 degree is probably too conservative; since there is a whole space of alternatives, not just one. But probably should not count on too many degrees of freedom.) Then we might call *any*  $f$  for which  $A_f/A_\phi < 20.09$  an "acceptable" distribution. If  $h=0$  is among them, then we may choose to report just the Gaussian intervals.

See further notes from KK and the papers in the book *Configural Polysampling* (S. Morgenthaler and J.W. Tukey 1990, Wiley.)

#### IV. Conclusions

- Point estimates without uncertainty are useless
- Uncertainty must incorporate not only random measurement error but also random shape (data distribution)
- Today, we can do this by taking advantage of computing power: considering different situations/techniques that before were not feasible.