Motivating Examples

Tests for exponentiality, uniformity, normality:

Reference: Lockhart, R. A. (1985) The asymptotic distribution of the correlation coefficient in testing fit to the exponential distribution. *The Canadian Journal of Statistics*, **13**, 253–256.

Lockhart, R. A. and M. A. Stephens (1998). The Probability Plot: Tests of Fit Based on the Correlation Coefficient. Chapter 16 in *Handbook of statistics, vol. 17. Order Statistics: Applications*. Eds: N. Balakrishnan, C. R. Rao. Elsevier: Amsterdam.

Why do we do all this mathematics?

Suppose $X_1, \ldots, X_n$ are an iid sample of real rvs.

Notation: F is distribution of individual X_i

Goal: test the hypothesis that F is an Exponential distbn unknown location and scale.

Means $(X_i - \alpha)/\beta$ is standard exponential.

Graphical technique: sort Xs

Get order statistics

$$X_{(1)} < \cdots < X_{(n)}$$

Notice

$$\mathrm{E}(X_{(i)}) = \alpha + \beta m_i$$

where m_i is expected value in standard case.

Plot $X_{(i)}$ against m_i.

Should be straight line.

Measure straightness somehow?

One statistic:

$$T_n = \frac{\text{Error Sum of Squares}}{\hat{\beta}^2} = n(1 - r^2)$$

where r is usual correlation coefficient.

Reject if T_n too big.

How big?

Power?

Make distributional approximation?

Techniques:

1. Structure of exponential order statistics
2. Parameters known first
3. Slutsky's theorem
4. Martingale central limit theorem
5. Contiguity

Theorem 1 : *Let* $V_1,\ldots,V_n$ *be iid standard exponentials. Put*

$$W_i = \alpha + \beta \sum_{j=1}^{i} V_j/(n+1-j)$$

Then $W_1,\ldots,W_n$ *have the same joint distribution as* $X_{(1)},\ldots,X_{(n)}$.

Proof: Compute joint densities and prove they are equal.

Shows $m_i = \sum_{j=1}^{i} 1/(n+1-j)$.

What estimates?

Standard least squares:

$$\begin{aligned}\hat{\beta} &= \sum_{j=1}^{n}(m_j-\bar{m})X_{(j)}/\sum_{j=1}^{n}(m_j-\bar{m})^2\\ \hat{\alpha} &= \bar{X}-\hat{\beta}\bar{m}\end{aligned}$$

Simplification: $\bar{m}=1$.

Alternatively: use MLEs.

Now plug in W_is to get expression in terms of Vs.

Fact: α and β disappear. The statistic is location scale invariant. So study statistic when $\alpha = 0$ and $\beta = 1$.

Do simpler case first. Don't estimate α, β. Statistic is

$$S_n \equiv \sum (X_{(i)} - m_i)^2$$

Rewrite as

$$S_n = \sum_{i=1}^{n} \sum_{j=}^{n} Q_{ij}(V_i - 1)(V_j - 1)$$

where

$$Q_{ij} = 1/\left\{n + 1 - \min(i, j)\right\}$$

Making distributional approximation. General principle: must stabilize mean and variance (or other location and scale measures).

First

$$\begin{aligned}\mu_n \equiv \mathsf{E}(S_n) &= \sum_{i=1}^{n} \mathsf{Var}(V_i) Q_{ii} \\ &= 2\sum_{i=1}^{n} 1/i \\ &= 2\log(n) + O(1)\end{aligned}$$

Goes to infinity so will have to consider

$$S_n - \mu_n$$

Second, put $W_{ij} = (V_i - 1)(V_j - 1)$:

$$\begin{aligned}\sigma_n^2 &\equiv \mathsf{Var}(S_n) \\ &= \sum_{ijkl} Q_{ij} Q_{kl} Cov\left\{W_{ij}, W_{kl}\right\}\end{aligned}$$

The ij, kl covariance is 0 unless the $\{i,j\} = \{k,l\}$.

This leaves the cases:

1. $i = k, j = l$ and $i \neq j$. The covariance is
$$\mathrm{E}\left\{(V_i - 1)^2 (V_k - 1)^2\right\} = 1$$

2. $i = l, j = k$ and $i \neq j$. Again the covariance is 1.

3. $i = j = k = l$. The covariance is
$$\mathrm{E}\left\{(V_i - 1)^4\right\} - \mathrm{E}^2\left\{(V_i - 1)^2\right\} = 8.$$

This makes

$$\sigma_n^2 = 2 \sum_{i \neq j} Q_{ij}^2 + 8 \sum_i Q_{ii}^2$$

You need to be able to develop approximations for this sort of sum.

Rethink: notice all those covariances which were 0.

Define

$$S_{1,n} = \sum_i Q_{ii}(V_i - 1)^2$$

and

$$S_{2,n} = \sum_{i \neq j} Q_{ij}(V_i - 1)(V_j - 1)$$

Let $\mu_{i,n}$ and $\sigma^2_{i,n}$ be the corresponding means and variances.

Note that

$$\begin{aligned} \mu_{2,n} &= 0 \\ \mu_{1,n} &= \mu_n \\ \sigma^2_{2,n} &= 2\sum_{i \neq j} Q^2_{ij} \\ \sigma^2_{1,n} &= 8\sum_i Q^2_{ii} \end{aligned}$$

so that $S_{1,n}$ and $S_{2,n}$ are uncorrelated.

Asymptotic formulas.

$$\sigma_{2,n}^2 = 8\sum_1^n 1/i^2 \to 8\pi^2/6.$$

$$\begin{aligned}\sigma_{1,n}^2 &= 4\sum_1^n (n-i)/(n+1-i)^2 \\ &= 4\log n + O(1)\end{aligned}$$

Off diagonal terms have larger SD so study

$$\frac{S_n - \mu_n}{\sigma_n}$$

Recognize that

$$\frac{S_{1,n} - \mu_n}{\sigma_n} \Rightarrow 0.$$

According to Slutsky's: if

$$\frac{S_{2,n}}{\sigma_n} \Rightarrow Z$$

for some rv Z then

$$\frac{S_n - \mu_n}{\sigma_n} \Rightarrow Z$$

Now how to do $S_{2,n}$?

Structure:

$$S_{2,n} = \sum_{i=2}^{n} A_i(V_i - 1).$$

Note A_i is function of $V_1, \ldots, V_{i-1}$ only. Namely:

$$A_i = 2\sum_{j=1}^{i-1} Q_{ij}(V_j - 1)$$

For any such structure the terms are uncorrelated!

So: if there is enough independence you might get normality.

In fact if we put

$$M_{k,n} = \sum_{i=2}^{k} A_i(V_i - 1)$$

then for each n

$$M_{2,n}, M_{3,n}, \cdots, M_{n,n}$$

is a martingale.

Use martingale central limit theorem.

Prove

$$\frac{S_{2,n}}{\sigma_n} \Rightarrow N(0,1)$$

Next: consider impact of estimation.

Write Error sum of squares, ESS, as function of parameter values:

$$ESS(\alpha,\beta) = \sum (X_{(i)} - \alpha - \beta m_i)^2$$

Compare ESS$(\hat\alpha, \hat\beta)$ to ESS(α_o, β_o).

Use Slutsky several times more.

Example: Now switch from Exponential to Uniform.

Consider again Error Sum of Squares.

(Ie: known parameters first.)

Notation: order statistics of sample are

$$U_{(1)} < \cdots < U_{(n)}$$

Structure: Same distribution as

$$\frac{V_1}{V_1 + \cdots + V_{n+1}}, \ldots \frac{V_1 + \cdots + V_n}{V_1 + \cdots + V_{n+1}}$$

for iid exponential Vs.

Error Sum of Squares is ratio N/D where

$$\begin{aligned} N &= \frac{1}{n}\sum_{k=1}^{n}\left(S_k - \frac{k}{n}S_{n+1}\right)^2 \\ S_k &= V + 1 + \cdots + V_k \\ D &= S_{n+1}^2/n \end{aligned}$$

Study behaviour of N because D is like n.

Imagine plotting successive terms in N against k or better, against k/n.

Think of k/n as variable t running from about 0 to about 1.

See factor $1/n$ as dt.

Formally define function

$$W_n(t) = n^{-1/2}(S_k - (k/n)S_{n+1})$$

where k is the integer part of nt. Then N is exactly

$$\int_0^1 W_n^2(t)dt$$

Weak convergence in $\mathcal{D}[0,1]$: tool to study such objects.

Conclusion

$$\text{ESS} \Rightarrow \sum_{i=1}^{\infty} \lambda_i Z_i^2$$

where Z_i are iid N(0,1) and λ are eigenvalues of integral equation.

Relevant tools: Hilbert space, compact operators on Hilbert space, Gaussian processes, differential equations.

Final version: what about testing for normality?

Statistics produce "asymptotically equivalent" to Shapiro-Wilk statistic.

Put $X_{(k)} = \Phi^{-1}(U_{(k)})$ to relate normal order statistics to uniform.

Use Taylor expansion (delta method) to relate $X_{(k)} - \mathsf{E}(X_{(k)})$ to $S_k - k/n$.

Do delicate analysis for k near 0 or n.

Next: develop relevant tools.