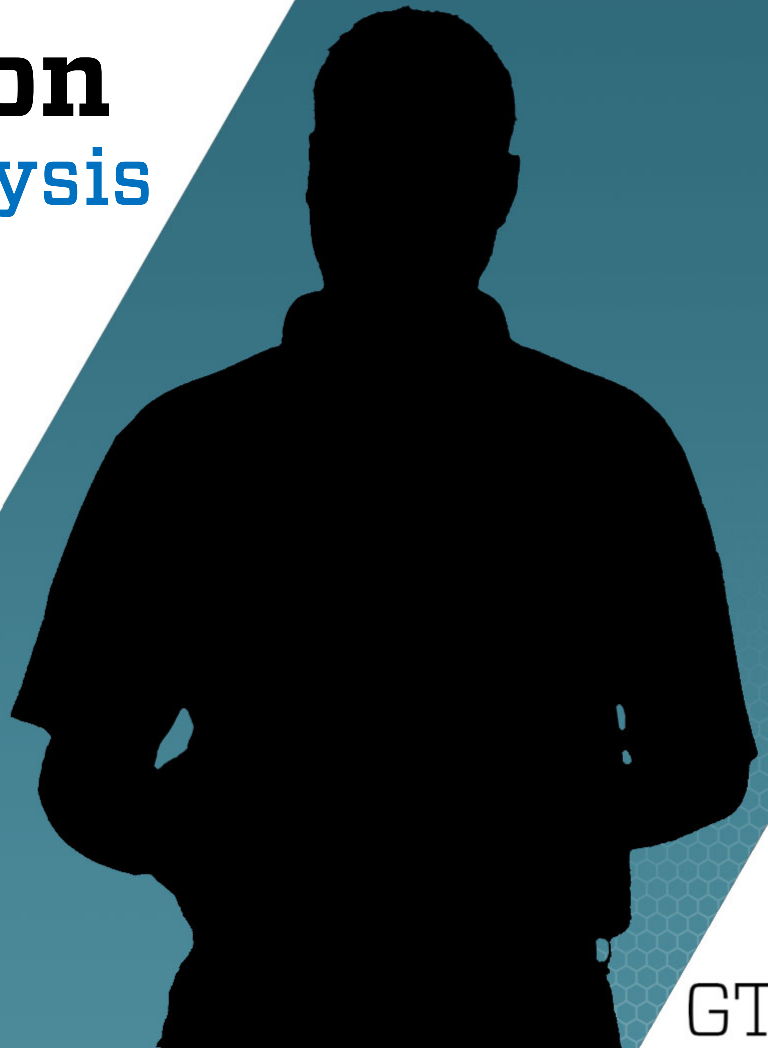
Computer Simulation

Module 8: Input Data Analysis

Dave Goldsman, Ph.D.
Professor
Stewart School of Industrial and Systems Engineering



Goodness-of-Fit Tests

Lesson Overview

Last Time: Finished off (for now) point estimation by discussing the Method of Moments.

This Time: We'll finally start our material on goodness-of-fit tests.

Idea: Does the choice of a particular simulation input distribution actually reflect reality?

At this point, we've guessed a reasonable distribution and then estimated the relevant parameters. Now let's conduct a formal test to see just how successful our toils have been.

In particular, test

$$H_0 : X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{p.m.f. / p.d.f. } f(x)$$

at level of significance

$$\alpha \;\equiv\; P(\text{Reject } H_0 | H_0 \text{ true}) \;=\; P(\text{Type I error})$$

(usually, $\alpha = 0.05$ or 0.01).

High-level view of a goodness-of-fit test procedure:

1. Divide the domain of $f(x)$ into k sets, say, $A_1, A_2, \ldots, A_k$ (distinct points if X is discrete or intervals if X is continuous).
2. Tally the actual number of observations that fall in each set, say, O_i, $i = 1, 2, \ldots, k$. If $p_i \equiv P(X \in A_i)$, then $O_i \sim \mathrm{Bin}(n, p_i)$.
3. Determine the expected number of observations that would fall in each set if H_0 were true, say, $E_i = \mathrm{E}[O_i] = np_i$, $i = 1, 2, \ldots, k$.
4. Calculate a test statistic based on the differences between the E_i's and O_i's. The chi-squared g-o-f test statistic is

$$\chi_0^2 \equiv \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$

5. A large value of χ_0^2 indicates a bad fit.

We *reject* H_0 if $\chi_0^2 > \chi^2_{\alpha,k-1-s}$, where

- s is the number of unknown parameters from $f(x)$ that have to be estimated. E.g., if $X \sim \text{Nor}(\mu, \sigma^2)$, then $s = 2$.
- $\chi^2_{\alpha,\nu}$ is the $(1-\alpha)$ quantile of the χ^2_ν distribution, i.e., $P(\chi^2_\nu < \chi^2_{\alpha,\nu}) = 1 - \alpha$.

If $\chi_0^2 \le \chi^2_{\alpha,k-1-s}$, we *fail to reject* H_0.

Remarks:

- Usual recommendation from baby stats class: For the χ^2 g-o-f test to work, pick k, n such that $E_i \geq 5$ and n at least 30.

- If the d.f. $\nu = k - 1 - s$ happens to be very big, then

$$\chi^2_{\alpha,\nu} \approx \nu \left[1 - \frac{2}{9\nu} + z_\alpha \sqrt{\frac{2}{9\nu}} \right]^3,$$

  where z_α is the appropriate standard normal quantile.

- Other g-o-f tests: Kolmogorov–Smirnov, Anderson–Darling, Shapiro–Wilk, etc.

Baby Example: Test H_0: X_i's are Unif(0,1). Suppose we have $n = 1000$ observations divided into $k = 5$ intervals.

interval	[0,0.2]	(0.2,0.4]	(0.4,0.6]	(0.6,0.8]	(0.8,1.0]
E_i	200	200	200	200	200
O_i	179	208	222	199	192

Turns out that $\chi_0^2 \equiv \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = 5.27$.

Let's take $\alpha = 0.05$. There are no unknown parameters, so $s = 0$. Then $\chi^2_{\alpha,k-1-s} = \chi^2_{0.05,4} = 9.49$.

Since $\chi_0^2 \leq \chi^2_{\alpha,k-1-s}$, we fail to reject H_0. $\square$

Discrete Example: The number of defects in printed circuit boards is hypothesized to follow a Geometric(p) distribution. A random sample of $n = 70$ printed boards has been collected, and the number of defects observed.

# defects	frequency
1	34
2	18
3	2
4	9
5	7
	70

We'll test H_0: The observations are Geom(p).

Start by estimating p via the MLE. The likelihood function is

$$L(p) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} (1-p)^{x_i - 1} p = (1-p)^{\sum_{i=1}^{n} x_i - n} p^n$$

$$\ell\text{n}(L(p)) = \left(\sum_{i=1}^{n} x_i - n \right) \ell\text{n}(1-p) + n\,\ell\text{n}(p)$$

$$\frac{d\,\ell\text{n}(L(p))}{dp} = \frac{-\sum_{i=1}^{n} x_i + n}{1-p} + \frac{n}{p} = 0.$$

Solving for p gives the MLE

$$\hat{p} = \frac{1}{\bar{X}} = \frac{70}{1(34)+2(18)+3(2)+4(9)+5(7)} = 0.476.$$

Let's get the g-o-f test statistic, χ_0^2. We'll make a little table, assuming $\hat{p} = 0.476$ is correct. By the Invariance Property of MLEs (this is why we learned it!), the expected number of boards having a certain value x is $E_x = nP(X = x) = n(1-\hat{p})^{x-1}\hat{p}$. Note that I'll combine the entries in the last row (≥ 5) so the probabilities add up to one.

x	$P(X = x)$	E_x	O_x
1	0.4762	33.33	34
2	0.2494	17.46	18
3	0.1307	9.15	2
4	0.0684	4.79	9
≥ 5	0.0752	5.27	7
	1.0000	70	70

Well, we really ought to combine the last two cells too, since $E_4 = 4.79 < 5$. Let's do so to get the following improved table.

x	$P(X = x)$	E_x	O_x
1	0.4762	33.33	34
2	0.2494	17.46	18
3	0.1307	9.15	2
≥ 4	0.1436	10.06	16
	1.0000	70	70

Thus, the test statistic is

$$\chi_0^2 = \sum_{x=1}^{4} \frac{(E_x - O_x)^2}{E_x} = \frac{(33.33 - 34)^2}{33.33} + \cdots = 9.12.$$

Let $k = 4$ denote the number of cells (that we ultimately ended up with), and let $s = 1$ denote the number of parameters we had to estimate.

Suppose the level $\alpha = 0.05$. Then we compare χ_0^2 against $\chi^2_{\alpha,k-1-s} = \chi^2_{0.05,2} = 5.99$.

Since $\chi_0^2 > \chi^2_{\alpha,k-1-s}$, we reject H_0. This means that the number of defects probably isn't geometric. $\square$

Summary

This Time: We had the fits – goodness-of-fits, that is! We did a high-level overview and then a couple of easy examples.

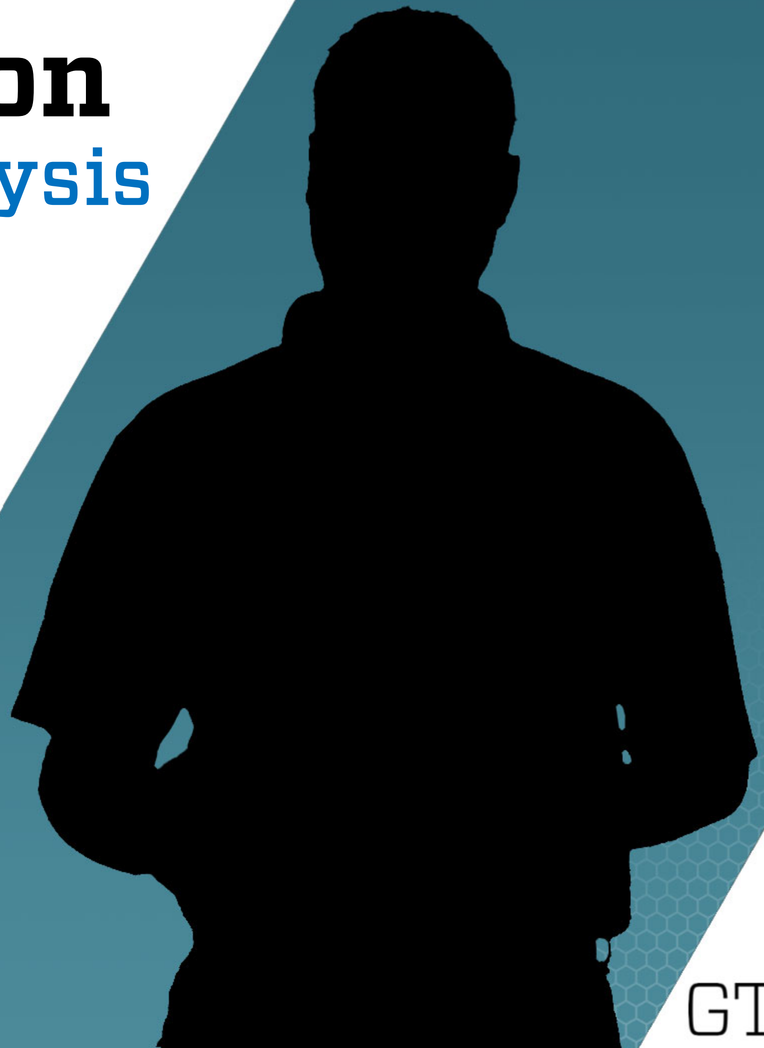
Next Time: We'll apply a g-o-f test to the exponential distribution.

GTx

Computer Simulation

Module 8: Input Data Analysis

Dave Goldsman, Ph.D.
Professor
Stewart School of Industrial and Systems Engineering



Exponential Example

GTx

Lesson Overview

Last Time: Had a high-level discussion of chi-square g-o-f tests and did some easy examples.

This Time: We'll do a g-o-f test for the exponential distribution.

The same general recipe can be applied to other distributions.

Continuous Distributions: For the continuous case, let's denote the intervals $A_i \equiv (a_{i-1}, a_i]$, $i = 1, 2, \ldots, k$. For convenience, we choose the a_i's to ensure that we have equal-probability intervals, i.e.,

$$p_i \;=\; P(X \in A_i) \;=\; P(a_{i-1} < X \leq a_i) \;=\; 1/k \quad \text{for all } i.$$

In this case, we immediately have $E_i = n/k$ for all i; and then

$$\chi_0^2 \;=\; \sum_{i=1}^{k} \frac{(O_i - (n/k))^2}{n/k}.$$

The issue is that the a_i's might depend on unknown parameters.

Example: Suppose that we're interested in fitting a distribution to a series of interarrival times. Could they be *Exponential*?

$$H_0 : X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exp}(\lambda).$$

Let's do a χ^2 g-o-f test with *equal-probability intervals*. This amounts to choosing a_i's such that

$$F(a_i) \;=\; P(X \leq a_i) \;=\; 1 - e^{-\lambda a_i} \;=\; \frac{i}{k}, \quad i = 1, 2, \ldots, k.$$

$$a_i \;=\; -\frac{1}{\lambda}\ell\text{n}\Big(1 - \frac{i}{k}\Big), \quad i = 1, 2, \ldots, k.$$

Great, but λ is unknown (so we'll estimate $s = 1$ parameter).

Good News: We know that the MLE is $\widehat{\lambda} = 1/\bar{X}$. Thus, by the Invariance Property, the MLEs of the a_i's are

$$\widehat{a}_i \;=\; -\frac{1}{\widehat{\lambda}}\,\ell\mathrm{n}\Big(1 - \frac{i}{k}\Big) \;=\; -\bar{X}\,\ell\mathrm{n}\Big(1 - \frac{i}{k}\Big), \quad i = 1, 2, \ldots, k.$$

Continue Example: Suppose that we take $n = 100$ observations and divide them into $k = 5$ equal-prob intervals. Further suppose that sample mean based on the 100 observations is $\bar{X} = 9.0$.

$$\widehat{a}_i \;=\; -9.0\,\ell\mathrm{n}\big(1 - 0.2i\big), \quad i = 1, \ldots, 5.$$

Suppose we determine which interval each of the 100 observations belongs to and tally them up to get the O_i's.

interval $(a_{i-1}, a_i]$	O_i	$E_i = \frac{n}{k}$
(0, 2.01]	25	20
(2.01, 4.60]	27	20
(4.60, 8.25]	23	20
(8.25, 14.48]	13	20
(14.48, ∞)	12	20
	100	100

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = 9.80$$

$$\chi^2_{\alpha,k-1-s} = \chi^2_{0.05,3} = 7.81$$

So we reject H_0. These observations ain't Expo. $\square$

Summary

This Time: We implemented a chi-square goodness-of-fit test for the exponential distribution… easy as pie!

Next Time: We'll do the same for a Weibull distribution. This takes a little more work than the Expo, but it's a more-robust g-o-f test.