Probability Distribution Modeling

Quiz Study Guide — Appendix B

Manuel D. Rossetti, PhD P.E.

How to Use This Deck

This deck is a **checklist**, not a summary. Every slide is phrased as something you should be able to *state*, *derive*, or *compute* without looking it up.

- **Question formats:** multiple choice, true/false, matching, fill-in-the-blank, multi-slot fill-in-the-blank, and numeric answers.
- Work the **Drills** slides by hand first; the answers are in the appendix at the end.
- The **Common Traps** slides are built from the actual wrong answers — read them last.

> What this appendix is about
>
> **Input modeling:** turning collected data into the probability distributions that drive a simulation — and testing whether the fit is defensible.

What This Appendix Covers

Fitting

- The input modeling process
- Discrete versus continuous
- Summarizing data
- Method of moments and maximum likelihood

Testing

- Chi-squared and Kolmogorov-Smirnov
- P-P and Q-Q plots, Anderson-Darling
- Testing pseudo-random numbers
- Choosing a distribution, and the cautions

Objectives — Fitting

You should be able to:

- Put the input-modeling steps in order and state each one's purpose
- State the principle for deciding discrete versus continuous
- Compute the mean, variance, coefficient of variation, and median
- Apply both histogram interval rules
- Distinguish method of moments from maximum likelihood, and produce the standard results

Objectives — Testing

You should be able to:

- Write the chi-squared statistic, its degrees of freedom, and the expected-count rule
- Write the K-S statistic and state its advantages and its caveat
- Distinguish a P-P plot from a Q-Q plot
- Set up each of the pseudo-random number tests and give its degrees of freedom
- Match distributions to modeling situations, and state the practical cautions

The Input Modeling Process

Table 1: The process **begins with documenting the process** — not with picking a distribution, running the model, or estimating before collecting

Step	Purpose
Document the process	Define the random variable — when the task starts, when it ends, what triggers it
Collect data	Develop a sampling plan, run a pilot, then collect
Graphical and statistical analysis	Histograms, time-series plots, autocorrelation plots, basic statistics
Hypothesize distributions	Propose candidates from physical reasoning plus the analysis
Goodness of fit	Estimate parameters, then test via chi-squared, K-S, P-P and Q-Q plots

Discrete versus Continuous

A **discrete** random variable takes values from a specified **countable** list; a **continuous** one takes any value in an interval or collection of intervals.

> **!** The modeling principle
>
> The decision starts from **WHAT** is being collected and **HOW** it is being collected — **not** from looking at a data file. Values may have decimal places and still be **discrete in concept**.

This appears as both a multiple-choice item and a true/false item.

Summarizing Data

$$\bar{X}(n) = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S^2(n) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$c_v = \frac{\sqrt{\mathrm{Var}[X]}}{E[X]}, \qquad \hat{c}_v = \frac{s}{\bar{x}}, \qquad \hat{\gamma}_1 = \frac{(1/n)\sum(X_i - \bar{X})^3}{[S^2]^{3/2}}$$

Median. For n **odd**, the $\left(\frac{n+1}{2}\right)$th order statistic. For n **even**, the average of the $\frac{n}{2}$th and $\left(\frac{n}{2}+1\right)$th.

Histogram intervals. Square-root rule $k = \sqrt{n}$; **Sturges' rule** $k = \lfloor 1 + \log_2 n \rfloor$.

Parameter Estimation

- **Method of moments** — equate **sample** moments (mean, variance, …) to the corresponding **population** moments of the hypothesized distribution and solve.
- **Maximum likelihood** — choose the parameters that **maximize the joint PDF/PMF of the sample**, equivalently the log-likelihood.

Poisson MLE: $\hat{\lambda} = \frac{1}{k}\sum_{i=1}^{k} n_i$ — the **sample average** of the interval counts.

Gamma method of moments, with $E[X] = \alpha\beta$ and $\mathrm{Var}[X] = \alpha\beta^2$:

$$\hat{\alpha} = \frac{\bar{X}^2}{S^2}, \qquad \hat{\beta} = \frac{S^2}{\bar{X}}$$

p-values

A p-value is the **smallest α level at which the observed test statistic is significant**; equivalently, the probability — assuming H_0 true — of observing a test statistic **at least as extreme** as the one observed.

It is not the probability that H_0 is true.

Decision rule: reject H_0 when p-value $\leq \alpha$. A small p-value means the observed outcome is rare under H_0, so H_0 is not a plausible explanation. A p-value of 0.001 is **strong evidence against** H_0.

Chi-Squared Goodness of Fit

$$\chi_0^2 = \sum_{j=1}^{k} \frac{(c_j - np_j)^2}{np_j}, \qquad \mathrm{df} = k - s - 1, \qquad np_j \geq 5$$

- c_j is the **observed** count, np_j the **expected** count under H_0.
- k is the number of classes and s the number of parameters **estimated from the data**.
- Choose classes so the **expected** count per class is **at least 5**. Small expected counts inflate the statistic, so consolidate those classes.

Computer-lab arrivals example. `xtabs(N ~ Week + Period + Day, data = p2)` gives a p-value of 0.7384, so we **do not reject** independence; the 160 observations are treated as independent and a Poisson is fit with $\hat{\lambda} = 8.275$ per 15-minute interval. After consolidation there are 9 classes with $s = 1$.

The Kolmogorov-Smirnov Test

$$D_n = \max(D_n^+, D_n^-), \quad D_n^+ = \max_i \left\{ \frac{i}{n} - \hat{F}(x_{(i)}) \right\}, \quad D_n^- = \max_i \left\{ \hat{F}(x_{(i)}) - \frac{i-1}{n} \right\}$$

the **largest vertical deviation** between the empirical and hypothesized CDFs.

Advantages over chi-squared: more powerful, usable on **smaller** samples, and **independent of the choice of class intervals**.

Caveat: when parameters are **estimated from the data**, the test becomes **conservative —** the actual Type I error is **less** than the specified α.

For $U(0,1)$, $F(x) = x$, so $D_n^+ = \max_i\{i/n - x_{(i)}\}$ and $D_n^- = \max_i\{x_{(i)} - (i-1)/n\}$. For $n \geq 35$, $D_{0.05} \approx 1.36/\sqrt{n}$.

Empirical CDFs and Diagnostic Plots

$$\tilde{F}_n(x_{(i)}) = \frac{i}{n}, \qquad \text{continuity correction: } \frac{i - 0.5}{n}, \qquad \text{Blom: } \frac{i - 0.375}{n + 0.25}$$

The qth quantile is $x_q = F^{-1}(q)$.

P-P plot

$\tilde{F}_n(x_{(i)})$ **versus** $\hat{F}(x_{(i)})$ — *probability against probability*.

Q-Q plot

$x_{(i)}$ **versus** $\hat{F}^{-1}(q_i)$ with $q_i = (i - 0.5)/n$ — *quantile against quantile*.

In **both**, a good fit is an approximately straight line of **slope 1** through the origin — a **45-degree** line.

Anderson-Darling detects **tail** differences and has **higher power** than K-S for many popular distributions; it is standard output in commercial fitting software.

Testing Pseudo-Random Numbers

Table 2: At 90% the band is $\pm 1.645/\sqrt{n}$; under white noise $\mathrm{Var}(r_k) \approx 1/n$

Test	What it does	df / critical value
Chi-squared for $U(0,1)$	Bin $(0,1)$ into k equal sub-intervals	$k-1$; a, b known so $s=0$
K-S for $U(0,1)$	Largest deviation from $F(x)=x$	$1.36/\sqrt{n}$ for $n \geq 35$
Serial (multi-d)	Counts in k^d hypercubes	$k^d - 1$
Runs up / down	**Sequential patterns** of "+" and "-"	—
Autocorrelation	r_k versus lag k	Band $\pm 1.96/\sqrt{n}$ at 95%

Choosing a Distribution — Discrete

Distribution	Typical situation
Bernoulli(p)	An independent trial with success probability p
Binomial(n, p)	Sum of n Bernoulli trials
Geometric(p)	Number of Bernoulli trials until the first success
Poisson(λ)	Counts of occurrences in an interval, area, or volume
Discrete uniform	Equally likely over a range or list of values

Choosing a Distribution — Continuous

Distribution	Typical situation
Uniform	Lack of data; everything equally likely on an interval
Exponential	Time to perform a task, time between failures, distance between defects
Weibull	Time to failure, time to complete a task
Triangular	A rough model given a minimum, a maximum, and a most-likely value
Lognormal	Time to perform a task; products of many other quantities

Exponential signatures. $c_v = 1$, so $\hat{c}_v$ near 1 is *consistent with* — not proof of — an exponential. And it is **memoryless**: $P\{X > \Delta t + t \mid X > t\} = P\{X > \Delta t\}$.

Why Not the Normal for Task Times?

The normal is defined over **all** the real numbers, **including negatives**, and “delay for a negative time” causes errors in a discrete-event simulation.

Alternatives: truncated normal, lognormal, gamma, Weibull, exponential.

The **lognormal** is a convenient task-time model because it is specified by mean and variance and is defined on the **positive reals**.

Practical Cautions

- **Input Analyzer "Square Error"** criterion: $\sum_j (h_j - \hat{p}_j)^2$ — observed relative frequencies versus theoretical interval probabilities. It **depends on the histogram interval specification**, so check the **sensitivity** of the ranking to the number of intervals.
- **Goodness-of-fit caveat:** with **little data** the tests tend **not to reject anything**; with **lots of data** they tend to **reject everything**.
- **Discrete empirical** is right for the truck-loading data ($n = 40$ package counts in $\{100, \dots, 150\}$) because of the **limited size and limited variety** of the data values, with proportions $\{0.05, 0.15, 0.225, 0.375, 0.125, 0.075\}$.
- **Do not simply reuse a small sample:** it does not cover all possible values — the observed max may be 56.11 when the true max is higher — and it makes parameter-driven sensitivity analysis difficult.
- **Correlated continuous data:** first verify the sampling plan; if real, model it explicitly (**NORTA**, autoregressive) or, with large data, take a random sub-sample to break the correlation.

The `fitdistrplus` Package

Construct	Role
`plotdist(data, discrete = ...)`	Plots the empirical PMF/CDF (or PDF/CDF)
`fitdist(data, "pois")`	Fits a discrete Poisson by maximum likelihood
`fitdist(data, "gamma")`	Fits a continuous gamma by maximum likelihood
`gofstat(fp)`	Chi-squared, K-S, Cramer-von Mises, Anderson-Darling, AIC, BIC
`plot(fp)`	Empirical versus theoretical density, CDF, P-P, Q-Q

Gamma task-time example. With $\bar{X} = 13.412$ and $S^2 = 50.44895$, both the chi-squared test ($p = 0.433$) and the K-S test ($p = 0.9444$) fail to reject the gamma fit, and the P-P and Q-Q plots are approximately linear.

Drills — Statistics

1. Compute the sample median of $\{3, 5, 7, 7, 38\}$.
2. Compute the sample median of $\{3, 5, 7, 7\}$.
3. For $n = 100$, how many intervals does the square-root rule recommend? Sturges' rule?
4. Gamma method of moments with $\bar{X} = 13.412$ and $S^2 = 50.44895$. Compute $\hat{\alpha}$ and $\hat{\beta}$ to 3 decimals.

Drills — Goodness of Fit

5. A chi-squared test uses $k = 6$ classes and $s = 2$ estimated parameters. How many degrees of freedom?
6. A chi-squared test of $U(0, 1)$ uses $k = 10$ classes with $a = 0$, $b = 1$ known. How many degrees of freedom?
7. The computer-lab Poisson example ends with 9 classes and one estimated parameter. How many degrees of freedom?
8. Compute $D_{0.05}$ for $n = 100$. If $D_n = 0.108$, what is the conclusion?
9. A serial chi-squared test uses $d = 2$ and $k = 4$. How many degrees of freedom?

Common Traps

- $c_v = \sigma/\mu$ — not μ/σ, and not σ^2/μ^2.
- Chi-squared df is $k - s - 1$. Only $U(0,1)$ testing with known a, b uses $k - 1$, because there $s = 0$.
- Expected counts per class must be **at least** 5, not at most 5.
- Reject when **p-value** $\leq \alpha$. A p-value is not $P\{H_0 \text{ true}\}$.
- The K-S test is **conservative** when parameters are estimated — Type I error is **less** than α, not more.
- **P-P plots probability against probability; Q-Q plots quantile against quantile.**
- Discrete versus continuous is decided by **what and how you collect**, not by decimal places in the file.
- The Square Error criterion **does** depend on the histogram intervals.

Appendix — Drill Answers (1 of 2)

1. $n = 5$ is odd, so the 3rd order statistic: $\mathbf{7}$.
2. $n = 4$ is even, so $(5 + 7)/2 = \mathbf{6}$.
3. $\sqrt{100} = \mathbf{10}$; $\lfloor 1 + \log_2 100 \rfloor = \lfloor 7.6439 \rfloor = \mathbf{7}$. Note the floor applies to the **whole** expression.
4. $\hat{\alpha} = 13.412^2/50.44895 = 179.8817/50.44895 = \mathbf{3.566}$; $\hat{\beta} = 50.44895/13.412 = \mathbf{3.761}$. Check: $3.566 \times 3.761 \approx 13.41 = \bar{X}$.

Appendix — Drill Answers (2 of 2)

5. $\mathrm{df} = 6 - 2 - 1 = \mathbf{3}$.
6. The parameters are **known**, so $s = 0$ and $\mathrm{df} = 10 - 0 - 1 = \mathbf{9}$.
7. $\mathrm{df} = 9 - 1 - 1 = \mathbf{7}$.
8. $D_{0.05} = 1.36/\sqrt{100} = \mathbf{0.136}$. Since $0.108 < 0.136$, we **do not reject** H_0 — the numbers are consistent with $U(0, 1)$.
9. $\mathrm{df} = k^d - 1 = 4^2 - 1 = \mathbf{15}$.