STAT 6350
Analysis of Lifetime Data

Probability Plotting
Purpose of Probability Plots

Probability plots are an important tool for analyzing data and have been particularly popular in the analysis of life data. Probability plots are used to:

- Access the adequacy of a particular distributional model.

- To detect multiple failure modes or mixture of different populations.

- Obtain graphical estimates of model parameters (e.g. by fitting a straight line through the points on a probability plot).

- Displaying the results of a parametric maximum likelihood fit along with the data.

- Obtain, by drawing a smooth curve through the points, a semiparametric estimate of failure probabilities and distributional quantiles.
Main Idea: For a given cdf, $F(t)$, one can linearize the \{ $t$ versus $F(t)$ \} plot by:

- Finding transformations of $F(t)$ and $t$ such that the relationship between the transformed variable is linear.

- The transformed axes can be relabeled in terms of the original probability and time variables.

The resulting probability axis is generally nonlinear and is called the probability scale. The data axis is usually a linear axis or a log axis.
Linearizing the Exponential CDF

- **CDF:** \( p = F(t; \theta, \gamma) = 1 - \exp\left[-\frac{(t-\gamma)}{\theta}\right], \ t \geq \gamma. \)

- **Quantile:** \( t_p = \gamma - \theta \log(1 - p). \)

**Conclusion:**

- The \( \{ t_p \text{ versus } -\log(1-p) \} \) plot is a straight line.

- Plot \( t_p \) on the horizontal axis (\( x \)-axis) and \( p \) on the vertical axis.

- \( \gamma \) is the **intercept** on the time axis and \( 1/\theta \) is equal to the **slope** of the cdf line (quantile vs. time).

- Changing \( \theta \) changes the slope of the line and changing \( \gamma \) changes the position of the line.
Linearizing the Normal CDF

- CDF: \[ p = F(y; \mu, \sigma) = \Phi_{nor} \left( \frac{y-\mu}{\sigma} \right), \quad -\infty < y < \infty. \]

- Quantile: \[ y_p = \mu + \sigma \Phi_{nor}^{-1}(p). \]
  where \( \Phi_{nor}^{-1} \) is the \( p \) quantile of the standard normal distribution.

Conclusion:

- The \{ \( y_p \) versus \( \Phi_{nor}^{-1}(p) \) \} plot is a straight line.

- \( \mu \) is the point at the time axis where the cdf intersects the \( \Phi^{-1}(p) = 0 \) line (i.e., \( p = 0.5 \)). The slope of the cdf line on the graph is \( 1/\sigma \).

- Any normal cdf plots as a straight line with positive slope. Also, any straight line with positive slope corresponds to a normal distribution.
Linearizing the Lognormal CDF

- **CDF:** \( p = F(t; \mu, \sigma) = \Phi_{\text{nor}} \left( \frac{\log(t) - \mu}{\sigma} \right), \ t > 0. \)

- **Quantile:** \( t_p = \exp \left[ \mu + \sigma \Phi_{\text{nor}}^{-1}(p) \right]. \)

Then \( \log(t_p) = \mu + \sigma \Phi_{\text{nor}}^{-1}(p). \)

**Conclusion:**

- The \( \{ \log(t_p) \text{ versus } \Phi_{\text{nor}}^{-1}(p) \} \) plot is a straight line.

- \( \exp(\mu) \) can be read from the time axis at the point where the cdf intersects the \( \Phi_{\text{nor}}^{-1}(p) = 0 \) line. The slope of the cdf line on the graph is \( 1/\sigma. \)

- Any given lognormal cdf plots as a straight line with positive slope. Also, any straight line with positive slope corresponds to a lognormal distribution.
Linearizing the Weibull CDF

- **CDF:**
  \[ p = F(t; \eta, \beta) = \Phi_{sev} \left( \frac{\log(t) - \mu}{\sigma} \right) = 1 - \exp \left[ -\left( \frac{t}{\eta} \right)^\beta \right], t > 0. \]

- **Quantile:**
  \[ t_p = \exp \left[ \mu + \Phi_{sev}^{-1}(p)\sigma \right] = \eta \left( -\log(1 - p) \right)^{1/\beta}. \]

- **This leads to**
  \[ \log(t_p) = \mu + \sigma \log\left[ -\log(1 - p) \right] = \log(\eta) + \log\left[ -\log(1 - p) \right]^{1/\beta}. \]
Conclusion:

- The \{ \log(t_p) \text{ versus } \log[-\log(1-p)] \} plot is a straight line.

- \( \eta = \exp(\mu) \) can be read from the time axis at the point where the cdf intersects the \( \log[-\log(1-p)] = 0 \) line, which corresponds to \( p \approx 0.632 \).

- The slope of the cdf line on the graph is \( \beta = 1/\sigma \) (but in the computations use base \( e \) logarithms for the times rather than the base 10 logarithms used for the figures).

- Any Weibull cdf plots as a straight line with positive slope. Also, any straight line with positive slope corresponds to a Weibull distribution.

- Exponential cdfs plot as straight lines with slopes equal to 1.
To construct a probability plot, one must decide how to plot the nonparametric estimate of $F(t)$ on the probability scales.

The discontinuity and randomness of $\hat{F}(t)$ make it difficult to choose a definition for pairs of points $(t, \hat{F})$ to plot.

When times reported as exact, it has been traditional to plot $\{ t_i \text{ versus } \hat{F}(t_i) \}$ at the observed failure times.

General idea: Plot an estimate of $F$ at some specified set of points in time and define plotting positions consisting of a corresponding estimate of $F$ at these points in time.
Criteria for Choosing Plotting Positions

Criteria for choosing plotting positions should depend on the application or purposes for constructing the probability plot.

Some applications that suggest criteria:

- *Checking Distributional Assumptions*

- *Estimation of Parameters*

- *Display of Maximum Likelihood Fits with Data*
Criteria for Choosing Plotting Positions

Checking Distributional Assumptions

- Probability plotting is used to check if the observed data are well approximated by the postulated parametric distribution.

- Some bias in the slope and location of the fitted line is not a serious problem.

- It is generally suggested that the choice of plotting positions, in moderated-to-large samples, is not so important.
Criteria for Choosing Plotting Positions

Estimation of Parameters

- Probability plotting is used to estimate parameters of a particular distribution based on a fitted line.

- The ‘best’ plotting positions will depend on the assumed underlying model and the functions to be estimated.

- For complete data, letting $i$ index the ordered observations, some general agreement on the plotting positions

$$p_i = \frac{i - 0.5}{n}.$$
Criteria for Choosing Plotting Positions

Display of Maximum Likelihood Fits with Data

- Probability plotting is used to display the ML fit and to compare with the corresponding nonparametric estimate.

- With a poor choice of plotting positions, the ML line may not fit the plotted points and the probability plot can give the false impression that the parametric model and the data disagree.

- Plotting simultaneous confidence bands on the probability plot will indicate the amount of sampling variability one might expect to see.
Choice of Plotting Positions

Three cases to consider:

- Continuous inspection (or small inspection intervals resulting in exact failures)

- Interval censored data with relative large intervals

- Arbitrarily censored data
Choice of Plotting Positions

Continuous Inspection Data and Single Censoring (or complete data)

- \( \hat{F}(t) \) is a step function increasing by an amount \( 1/n \) at each reported failure time until the last reported failure time.

- A reasonable compromise plotting position is the midpoint of the jump

\[
\frac{i - 0.5}{n} = \frac{1}{2} \left[ \hat{F}(t(i)) + \hat{F}(t(i-1)) \right].
\]

- Plotting Positions: \( \{ t(i) \text{ versus } \frac{i - 0.5}{n} \} \).
Choice of Plotting Positions

Continuous Inspection Data and Single Censoring (or complete data)

• Justification:

  – The median of the $i$-th order statistic in a sample of size $n$ is approximately $F^{-1}[(i - 0.5)/n]$.

  – Plotting at the bottom/top of the step would lead to bias in the plotted points and the ML line would tend to be above/below the plotted points.

  – If the last reported time is a failure, it is not possible to plot a point at $\hat{F}(t) = 1$. 
Choice of Plotting Positions

Continuous Inspection Data and Multiple Censoring

- $\hat{F}(t)$ is a step function until the last reported failure time but the step increases may be different than $1/n$

- Plotting Positions: \{ $t_{(i)}$ versus $p_i$ \} with

\[
p_i = \frac{1}{2} \left\{ \hat{F}[t_{(i)}] + \hat{F}[t_{(i-1)}] \right\}.
\]

- Justification: This is consistent with the definition for single censoring.
Choice of Plotting Positions

Interval-Censored Inspection Data

- Let the inspection times be \((t_0, t_1], (t_1, t_2], (t_{m-1}, t_m]\).

- The upper endpoints of the inspection intervals \(t_i, \ i = 1, 2, \ldots\), are convenient plotting times.

- Plotting Positions: \(\{ t_{(i)} \text{ versus } p_i \} \) with
  \[
p_i = \hat{F}[t_{(i)}].
  \]

- When there are no censored observations beyond \(t_m\), \(F(t_m) = 1\) and this point cannot be plotted on the probability paper.

- Justification: With no losses, from standard binomial theory,
  \[
  E[\hat{F}(t_i)] = F(t_i).
  \]
Extensions of Probability Plots

The probability plotting techniques can be extended to construct probability plots for:

- Distributions that are not members of the location-scale family.

- The help identify, graphically, the need for non-zero threshold parameter.

- Estimate graphically a shape parameter.
Distribution with a Threshold Parameter

- The lognormal, Weibull, gamma, and other similar distributions can be generalized by the addition of a threshold parameter, $\gamma$, to shift the beginning of the distribution away from 0.

- These distributions are particularly useful for fitting skewed distributions that are shifted far to the right of 0.

- For example, the cdf and quantiles of the three-parameter lognormal distribution can be expressed as

$$p = F(t; \mu, \sigma, \gamma) = \Phi_{\text{nor}} \left[ \frac{\log(t - \gamma) - \mu}{\sigma} \right], t > \gamma.$$
Distributions with one or more unknown shape parameters, the plotting scales depend on the given value or estimate for the shape parameter.

There are two approaches to specifying an unknown shape parameter for a probability plot:

- Plot the data with different given values of the shape parameter in an attempt to find a value that will give a probability plot that is nearly linear.

- Use parametric ML methods to estimate the shape parameter and use the estimate value to construct probability plotting scales.

These two approaches generally lead to approximately the same plot.
Linearizing the 3-Parameter Gamma CDF

- CDF: $p = F(t; \theta, \kappa, \gamma) = \Gamma_I\left(\frac{t - \gamma}{\theta}; \kappa\right)$, $t > \gamma$.

- Quantile: $t_p = \gamma + \theta \Gamma_I^{-1}(p; \kappa)$
  where $\Gamma_I(z; \kappa) = \int_0^z x^{\kappa-1}e^{-x}dx/\Gamma(\kappa)$ and $\Gamma(\kappa) = \int_0^\infty x^{\kappa-1}e^{-x}dx$.

Conclusion:

- $\{t_p$ versus $\Gamma_I^{-1}(p; \kappa)\}$ plot is a straight line.

- The probability axis depends on specification of the shape parameter $\kappa$.

- $\gamma$ is the intercept on the time axis (because $\Gamma_I^{-1}(p; \kappa) = 0$ when $p = 0$). The slope of the cdf line is equal to $1/\theta$.

- Changing $\theta$ changes the slope of the line and changing $\gamma$ changes the position of the line.
Linearizing the 3-Parameter Weibull CDF

• CDF: 
\[ p = F(t; \mu, \sigma, \gamma) = \Phi_{sev}\left[\frac{\log(t-\gamma)-\mu}{\sigma}\right], \quad t > \gamma. \]

• Quantile: 
\[ t_p = \gamma + \eta \left[-\log(1-p)\right]^{1/\beta} \]
where \( \Phi_{sev}(z) = 1 - \exp[-\exp(z)] \), \( \eta = \exp(\mu) \), \( \beta = 1/\sigma. \)

Conclusion:

• The \{ \( t_p \) versus \( [-\log(1-p)]^{1/\beta} \) \} plot is a straight line.

• The probability axis for this linear-time-axis Weibull probability plot requires specification of the shape parameter \( \beta \).

• \( \gamma \) is the intercept on the time axis. The slope of the cdf line is equal to \( 1/\eta \).

• The plot allows graphical estimation the threshold parameter \( \gamma \).
Linearizing the Generalized Gamma CDF

- CDF: \( p = F(t; \theta, \beta, \kappa) = \Gamma_I \left( \left( \frac{t}{\theta} \right)^\beta ; \kappa \right) \).

- Quantile: \( t_p = \theta \left[ \Gamma_I^{-1}(p; \kappa) \right]^{1/\beta} \)
  
  Then \( \log(t_p) = \log(\theta) + \frac{1}{\beta} \log \left[ \Gamma_I^{-1}(p; \kappa) \right]. \)

Conclusion:

- The \( \{ t_p \text{ versus } \log \left[ \Gamma_I^{-1}(p; \kappa) \right] \} \) plot is a straight line.

- The scale parameter \( \theta \) is the intercept on the time scale, corresponding to the time where the cdf crosses the horizontal line at \( \log \left[ \gamma_I^{-1}(p; \kappa) \right] = 0. \)

- The slope of the line on the graph with time on the horizontal axis is \( \beta. \)

- The probability scale for the GENG probability plot requires a given value of the shape parameter \( \kappa. \)
Application of Probability Plotting

• Using simulation to help interpret probability plots
  – Try different assumed distributions and compare the results.
  – Assess linearity, allowing for more variability in the tails.
    * Use simultaneous nonparametric confidence bands.
    * Use simulation or bootstrap to calibrate.

• Possible reasons for a bend in a probability plot
  – Shape bend or change in slope generally indicate an abrupt change in a failure process.
  – Causes for such behavior could include two or more failure modes or a mixture of different subpopulations.