Parameter Estimation

Statistical Inference

- One of the major applications of statistics is estimating population parameters from sample statistics.
- Methods of making inferences about parameters:
  - ➢**Estimating the parameter**
  - ➢**Testing a hypothesis** about the value of the parameter.
- Parameter Estimation is the pre-step for hypothesis test.

Statistical Inference

- One of the major applications of statistics is estimating population parameters from sample statistics.
- Methods of making inferences about parameters:
  - ➢**Estimating the parameter**
  - ➢**Testing a hypothesis** about the value of the parameter.
- Parameter Estimation is the pre-step for hypothesis test.

Parameter Estimation

- Parameters are descriptive measures of an entire population.
- Sample statistics is to **sample from the population to obtain parameter estimates.** One goal of statistical analyses is to obtain estimates of the population parameters along with the amount of error associated with these estimates
- There are several types of parameter estimates:
  - ❖Point estimates
  - ❖Confidence intervals

Point Estimation

- Point estimates are the single, most likely value of a parameter. For example, the point estimate of population mean (the parameter) is the sample mean (the parameter estimate).

- For example, a poll may seek to estimate the proportion of adult residents of a city that support a proposition to build a new sports stadium. Out of a random sample of 200 people, 106 say they support the proposition. Thus in the sample, 0.53 (106/200=0.53)of the people supported the proposition. This value of 0.53 is called a point estimate of the population proportion. It is called a point estimate because the estimate consists of a single value or point.

Confidence intervals

- Confidence intervals are a range of values likely to contain the population parameter.

- An example of a 95% confidence interval is shown below:

$$72.85 < \mu < 107.15$$

There is good reason to believe that the population mean lies between these two bounds of 72.85 and 107.15 since 95% of the time confidence intervals contain the true mean.

Confidence Interval Vs. Point Estimation

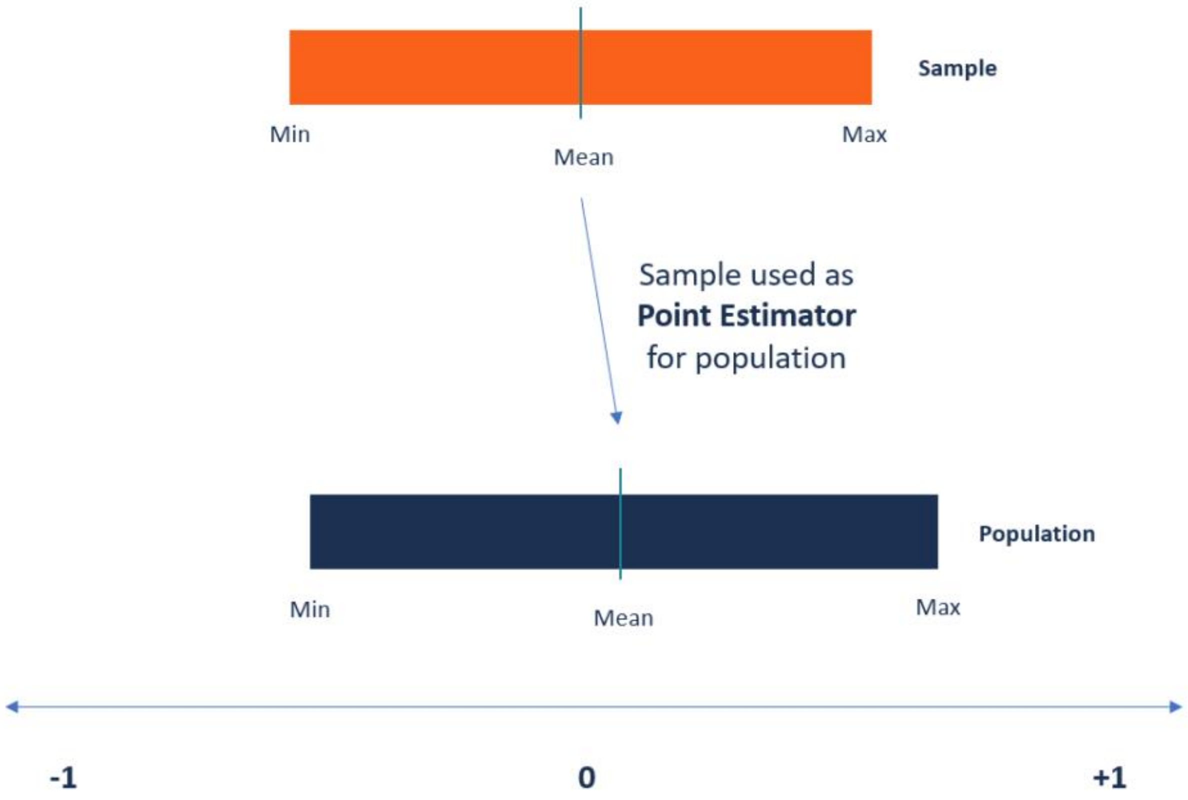
- Sample statistics are not expected to predict with absolute certainty the parameters of a population because of sampling error.
- If we used a single sample mean to estimate the mean of the population, we would be using a **point estimate** – a single value based on sample data to estimate the parameter of a population.
- However, the accuracy of our estimate would remain unknown.
- We could be more certain of our estimate by using a **confidence interval** – a range of values which, with a specified degree of confidence, is expected to include a population parameter.

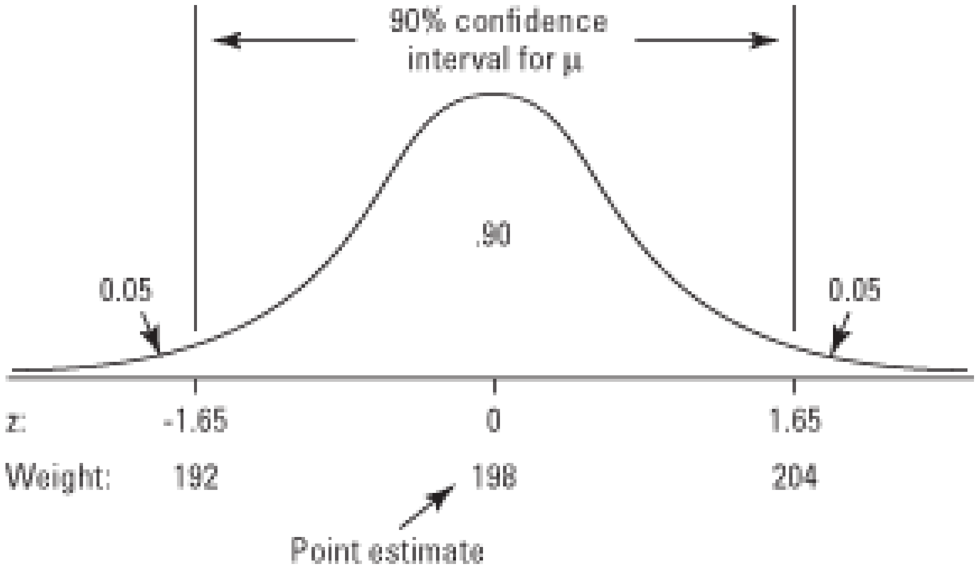
Point Estimates

For point estimates, the obtained sample values with no adjustments for standard error are reported (i.e., the numerator in the working formula for the t-statistic).

- For one-sample t-tests, the point estimate is: M
- For two-sample t-tests, independent measures design, the point estimate is: $M_1 – M_2$
- For two-sample t-tests, repeated measures design, the point estimate is: M_D

While point estimates are straightforward and readily understood, their values can only be expected to be "somewhere in the neighborhood" of the population values. Confidence intervals specify the boundaries of the neighborhood.


Sample
Min
Mean
Max
Sample used as
Point Estimator
for population
Population
Min
Mean
Max
-1
0
+1


90% confidence
interval for μ
.90
0.05
0.05
z:
-1.65
0
1.65
Weight:
192
198
204
Point estimate

Sample Bias

- Various types of sampling: e.g., *simple random sampling*, *stratified random sampling*

- Sampling bias refers to the method of sampling, not the sample itself.

- Research design or data collection method can lead to sampling bias. Sampling bias can occur in both probability and non-probability sampling.

Types of sampling bias

By David M. Lane

Type	Explanation	Example
Self-selection	People with specific characteristics are more likely to agree to take part in a study than others.	People who are more thrill-seeking are likely to take part in pain research studies. This may skew the data.
Undercoverage	Some members of a population are inadequately represented in the sample.	Administering general national surveys online may miss groups with limited internet access, such as the elderly and lower-income households.
Survivorship	Successful observations, people and objects are more likely to be represented in the sample than unsuccessful ones.	In scientific journals, there is strong publication bias towards positive results. Successful research outcomes are published far more often than null findings.

Sample Mean

- In any situation where we observe a simple random sample $X_1, X_2, \ldots, X_n$ from some population with mean μ, we know that the sample mean $\bar{X} = (X_1 + X_2 + \cdots + X_n)/n$ satisfies $E(\bar{X}) = \mu$, so it is natural to estimate μ by $\bar{X}$. We treat the Rasmussen survey as a binomial experiment with $E(\bar{X}) = p$, so using $\hat{p}$ $(= \bar{X})$ to estimate p is a special case of using $\bar{X}$ to estimate μ.

Estimator and Estimate

- It is important to distinguish between the rule that we follow to estimate a parameter and the value that we find for a particular sample.
- We call the rule an **estimator** and the value an estimate. For example, in the survey data, the rule is "estimate p by the sample fraction $\hat{p}$ ", and the value is .46. So the estimator is $\hat{p}$, and the estimate is .46. The same estimator $\hat{p}$ with a different sample gave a different estimate

Sampling Distribution

- Clearly a point estimator is a statistic, and therefore has a sampling distribution. Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from some population with a parameter θ, and that is a statistic that we want to use as an estimator of θ.

$$\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$$

Bias

- If E($\hat{\theta}$) = θ for all possible values of θ, $\hat{\theta}$ is an unbiased estimator of θ. In general, the bias of $\hat{\theta}$ as an estimator of θ is E($\hat{\theta}$ − θ) = E($\hat{\theta}$) − θ. A biased estimator in a sense systematically over-estimates or under-estimates θ, so we try to avoid estimators with large biases.
- An unbiased estimator is desirable, but not always available, and not always sensible.

- For example, suppose that n = 1, and X = X1 has the Poisson distribution with parameter μ,

$$P(X = x) = p(x; \mu) = e^{-\mu} \frac{\mu^x}{x!}, x = 0, 1, \ldots$$

- E(X) = μ, so X is an unbiased estimator of μ, but suppose that the parameter of interest is $\theta = e^{-\mu}$. The only unbiased estimator of θ is

$$\hat{\theta} = \begin{cases} 1 & \text{if } X = 0, \\ 0 & \text{if } X > 0. \end{cases}$$

Sampling Variability

- The sampling variability of a statistic refers to how much the statistic varies from sample to sample and is usually measured by its *standard error* ;

- the smaller the standard error, the less the sampling variability. For example, the *standard error of the mean* is a measure of the sampling variability of the mean. Recall that the formula for the standard error of the mean is

$$\sigma_M = \frac{\sigma}{\sqrt{N}}$$

The larger the sample size (N), the smaller the standard error of the mean and therefore the lower the sampling variability.

Mean Squared Error

- We measure how far an estimator $\hat{\theta}$ is from the true value θ using the mean squared error: MSE($\hat{\theta}$; θ) = E[($\hat{\theta}$ − θ) ^{2}]. We can show that MSE($\hat{\theta}$; θ) = (bias)^2 + V($\hat{\theta}$). For an unbiased estimator, bias = 0, so MSE($\hat{\theta}$; θ) = V($\hat{\theta}$)

Bias: $E(\hat{\theta} - \theta) = E(\hat{\theta}) - \theta$

- Many biased estimators are approximately unbiased, in the sense that

$$(\text{bias})^2 \ll V(\hat{\theta}),$$

$$\text{MSE}(\hat{\theta};\theta) \approx V(\hat{\theta}).$$

- Standard error: if an estimator is unbiased, either exactly or approximately, its performance is measured by $V(\hat{\theta})$, or by its standard deviation $\sigma_{\hat{\theta}}$ = sqrt($V(\hat{\theta})$), also known as its standard error.

$$\sigma_{\hat{\theta}} = \sqrt{V(\hat{\theta})},$$

- Often an estimator's standard error is a function of θ or other parameters; these must be replaced by estimates before we can actually calculate a value.

- Example: binomial distribution;

Example: binomial distribution; $V(\hat{p}) = p(1-p)/n$, so

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}, \text{ and } \hat{\sigma}_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Methods of Point Estimation

- In some situations we have an obvious estimator $\hat{\theta}$, such as the binomial $\hat{p}$ = X/n.
- In other cases we may not.

For example: Ozone pollution Suppose that $X_1, X_2, \ldots, X_{28}$ are daily maximum ozone levels on 28 consecutive days. Suppose further that we want to model these as independent variables with the Weibull distribution $f(x; \alpha, \beta) = \frac{\alpha}{\beta}\left(\frac{X}{\beta}\right)^{\alpha-1} e^{-\left(X/\beta\right)^{\alpha}}$, $0 < x < \infty$

- It is not obvious how to estimate either α or β. Suppose that we know from other data that α is well approximated by the value 2. It is still not obvious how to estimate β. Before we observed the data, the joint pdf

$$\prod_{i=1}^{n} f(x_i; \alpha, \beta)$$

- Likelihood function: After observing $x_1, x_2, \ldots, x_n$, we can use the same function to measure the relative likelihood of different values of α and β (or just β if we believe we know the value of $\alpha = \alpha_0$). When used this way, we call it the likelihood function

$$L(\beta) = \prod_{i=1}^{n} f(x_i; \alpha_0, \beta).$$

Methods of Point Estimation

- **Probability Plotting**: A method of finding parameter values where the data is plotted on special plotting paper and parameters are derived from the visual plot
- **Rank Regression (Least Squares)**: A method of finding parameter values that minimizes the sum of the squares of the residuals.
- **Maximum Likelihood Estimation:** A method of finding parameter values that, given a set of observations, will maximize the likelihood function.
- **Bayesian Estimation Methods**: A family of estimation methods that tries to minimize the posterior expectation of what is called the utility function. In practice, what this means is that existing knowledge about a situation is formulated, data is gathered, and then posterior knowledge is used to update our beliefs.

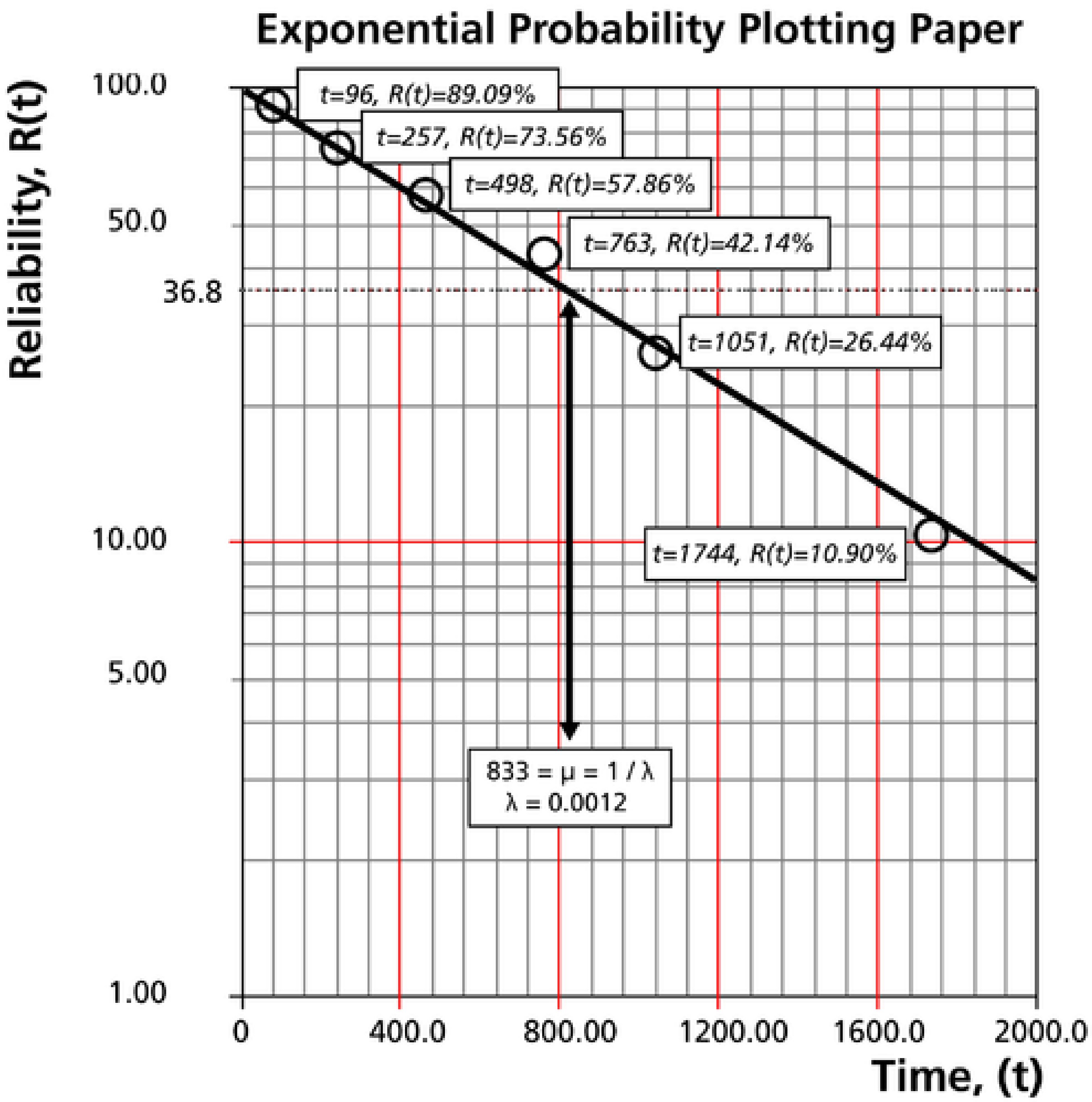
Qualities of Estimators

- If the value an estimator estimates for the parameter, θ', always converges to the actual parameter value θ as the quantity of data used for parameter estimation increases, we say an estimator is **consistent**.
- The bias of an estimator is the deviation of the expectation from the actual true value. If, for a given estimator, the bias is zero, we say that that estimator is **unbiased**.
- A third statistic that tells us about the reliability of an estimator is the variance. If an estimator has lower variance than another we say it is more **efficient**, and we can calculate the **efficiency** of estimator *p* relative to estimator *q* as $(\mathrm{Var}(\theta'_p))/\mathrm{Var}(\theta'_q)$.

Probability Plotting

- The least mathematically intensive method for parameter estimation is the method of probability plotting.
- The method of probability plotting takes the *cdf* of the distribution and attempts to linearize it by employing a specially constructed paper.
- The steps using the 2-parameter Weibull distribution as an example. This includes:
  - ➢ Linearize the unreliability function
  - ➢ Construct the probability plotting paper
  - ➢ Determine the X and Y positions of the plot points
  - ➢ And then using the plot to read any particular time or reliability/unreliability value of interest.

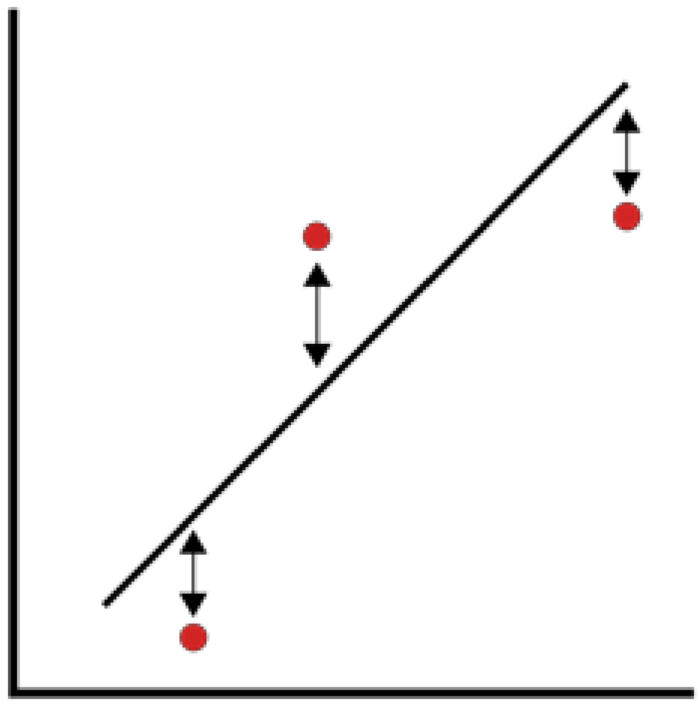
PDF	$f(x) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & x \geq 0 \\ 0 & x < 0 \end{cases}$
CDF	$\begin{cases} 1 - e^{-(x/\lambda)^k} & x \geq 0 \\ 0 & x < 0 \end{cases}$



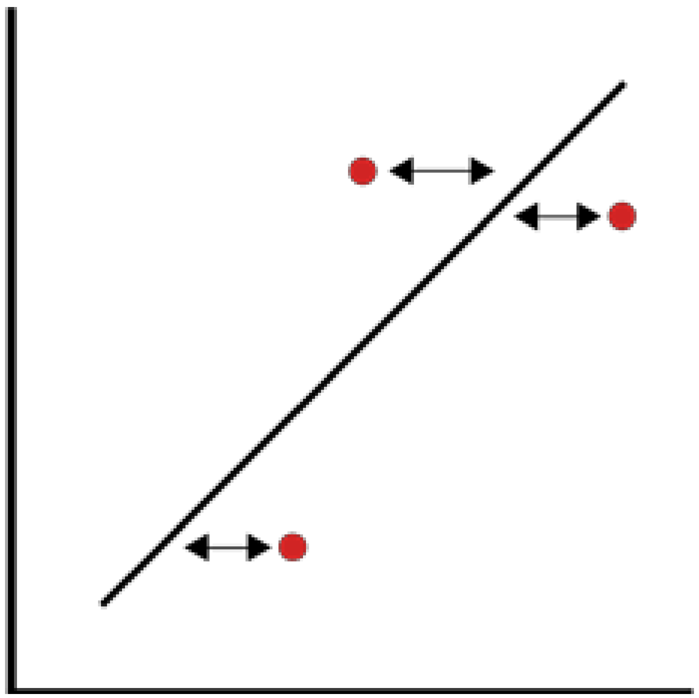
Exponential Probability Plotting Paper
Reliability, R(t)
100.0
50.0
36.8
10.00
5.00
1.00
t=96, R(t)=89.09%
t=257, R(t)=73.56%
t=498, R(t)=57.86%
t=763, R(t)=42.14%
t=1051, R(t)=26.44%
t=1744, R(t)=10.90%
833 = μ = 1 / λ
λ = 0.0012
0
400.0
800.00
1200.00
1600.0
2000.0
Time, (t)

Least Squares (Rank Regression)

- Regression analysis mathematically fits the best straight line to a set of points, in an attempt to estimate the parameters.
- The term rank regression is used instead of linear regression because the regression is performed on the rank values, more specifically, the median rank values (represented on the y-axis).
- The method of least squares requires that a straight line be fitted to a set of data points, such that the sum of the squares of the distance of the points to the fitted line is minimized. This minimization can be performed in either the vertical or horizontal direction.



Minimizing distance in the y-direction



Minimizing distance in the x-direction

Maximum Likelihood Estimation:

- Suppose we have a random sample $X_1,X_2,\cdots,X_n$ whose assumed probability distribution depends on some unknown parameter θ. Our primary goal here will be to find a point estimator $u(X_1,X_2,\cdots,X_n)$, such that $u(x_1,x_2,\cdots,x_n)$ is a "good" point estimate of θ where $x_1,x_2,\cdots,x_n$ are the observed values of the random sample.

- For example, if we plan to take a random sample $X_1, X_2, \cdots, X_n$ for which the X_i are assumed to be normally distributed with mean μ and variance σ^2, then our goal will be to find a good estimate of μ, say, using the data $x_1, x_2, \cdots, x_n$ that we obtained from our specific random sample.

- Suppose we have a random sample $X_1, X_2, \cdots, X_n$ for which the probability density (or mass) function of each X_i is $f(x_i;\theta)$. Then, the joint probability mass (or density) function of $X_1, X_2, \cdots, X_n$, which we'll (not so arbitrarily) call $L(\theta)$ is:

$$L(\theta) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = f(x_1;\theta) \cdot f(x_2;\theta) \cdots f(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

- The first equality is of course just the definition of the joint probability mass function. The second equality comes from that fact that we have a random sample, which implies by definition that the Xi are independent. And, the last equality just uses the shorthand mathematical notation of a product of indexed terms. Now, in light of the basic idea of maximum likelihood estimation, one reasonable way to proceed is to treat the "**likelihood function**" $L(\theta)$ as a function of θ, and find the value of θ that maximizes it.

Bayesian Parameter Estimation Methods

$$f(\theta|Data) = \frac{L(Data|\theta)\varphi(\theta)}{\int_{\zeta} L(Data|\theta)\varphi(\theta)d(\theta)}$$

where:

- θ is a vector of the parameters of the chosen distribution
- ζ is the range of θ
- $L(Data|\theta)$ is the likelihood function based on the chosen distribution and data
- $\varphi(\theta)$ is the prior distribution for each of the parameters

Properties of 'Good' Estimators

- The center of the sampling distribution for the estimate is the same as that of the population. When this property is true, the estimate is said to be **unbiased**. The most often-used measure of the center is the mean.
- The estimate has **the smallest standard error** when compared to other estimators.

For example, in the normal distribution, the mean and median are essentially the same. However, the standard error of the median is about 1.25 times that of the standard error of the mean. We know the standard error of the mean is σ_n. Therefore in a normal distribution, the SE(median) is about 1.25 times σ_n. This is why the mean is a better estimator than the median when the data is normal (or approximately normal).