An Introduction to Statistical Evaluation of Drug Products

by

Jen-pei Liu, Ph.D.

National Taiwan University and

National Health Research Institutes

jpliu@ntu.edu.tw

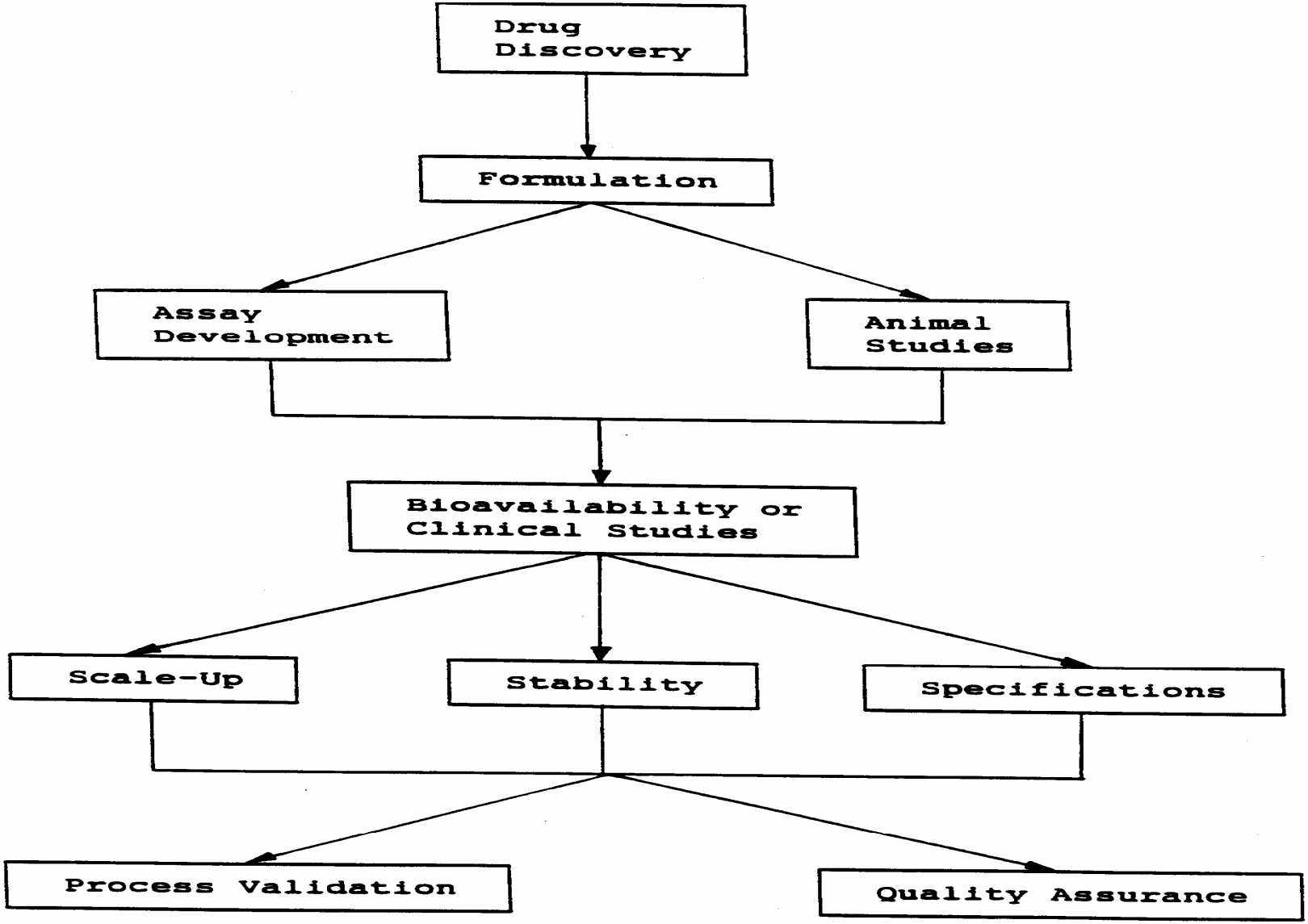
at

Feng-Chia University

November 17, 2005

Taichung, Taiwan

Drug Development



Outlines

Introduction

Effectiveness of Drug Products

Equivalence of Generic Drug Products

Estimation of Shelf-life of Drug Products

Quality Control of Drug Products

Evaluation of Diagnostic Devices

Summary

Introduction

- Evidence from clinical trials must prove that the drug is efficacious – drug is better than no drug
- Inference from the sample (patients in trials) to the targeted population (patients in clinical practice)
- A decision process for clinical hypotheses based on the trial objectives through statistical testing procedures

Introduction - Clinical Trials?

- FDA (21 CFR 312.3, April 1994)

  A clinical trial is the clinical investigation of a drug which is administered or dispensed to, or used involving one or more human subjects.

- Chow and Liu (2004)

  A clinical trial is the clinical investigation in which treatments are administered, dispensed or used involving one or more human subjects for evaluation of the treatments.

Introduction – Three Key Components

- Experimental unit

  A subject from a targeted population under study. For example

  - Healthy human subjects
  - Patients with certain diseases at certain stages

Introduction – Three Key Components

- Treatment

  It could be a placebo or any combinations of

  - A new pharmaceutical entity
  - A new diet
  - A surgical procedure
  - A diagnostic test
  - A medical device

Introduction – Three Key Components

- Evaluation

--Efficacy analysis

  - Clinical endpoints

--Safety assessment

  - Adverse experience
  - Laboratory test results

--Quality of life assessment

--Pharmacoeconomics analysis

--Outcomes research

Introduction – Statistical Designs

- Parallel Group Designs

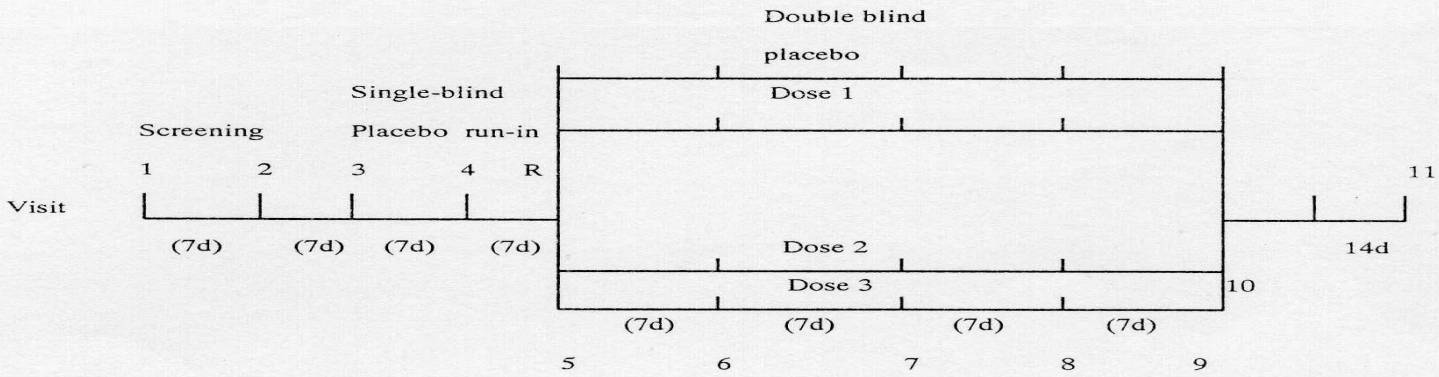
  The patients are randomized to one of two or more groups, each group being allocated to a different treatment.

  - Advantages
    - Simple and easy to implement.
    - Less complicated analysis and interpretation.
  - Drawbacks
    - Relative large variability

      Inter-patient + Intra-patient

Display 3b

STUDY DESIGN AND SCHEDULE OF ASSESSMENTS



Assessment	Screening	Run-in	Baseline	Treatment				Follow-up		
Study Week	-2	-1	0	1	2	3	4	5	6	8
Informed Consent	X									
History	X									
Physical Exam.	X									X
Effectiveness:										
primary variable	X	X	X	X	X	X	X	X	X	X
secondary variable	X	X	X	X		X			X	X
Safety:										
Adverse events	X	X	X	X	X	X	X	X	X	X
Lab. tests X		X	X			X		X	X	
Body weight	X		X						X	X

Effectiveness of Drug Products

- Example: Farlow et al (JAMA 1992; 268: 2523-2529)
- Randomized, double-blind, parallel groups
- Objective

  To compare the tacrine (20, 40, 80 mg per day) versus placebo for probable Alzheimer's disease
- Null hypothesis

  No difference in ADAS-cog scale between 80 mg of tacrine and placebo.
- Alternative hypothesis

  There exists a true difference in ADAS-cog scale between 80 mg of tacrine and placebo.

Effectiveness of Drug Products

- Example: The NINDS rt-PA Stroke Study Group (NEJM 1996; 335: 841-7)
- Objective for part I

  A greater proportion of patients with acute ischemic stroke treated with t-PA, as compared with those given placebo, have early improvement (>= 4 from baseline on NIHSS).
- Primary efficacy endpoint

  Proportion of patients with improvement
- Null hypothesis

  No difference in the proportions of patients with improvement between t-PA and placebo.
- Alternative hypothesis

  The minimal difference in the proportions of patients with improvement between t-PA and placebo is at least 24%.

Effectiveness of Drug Products

Statistical Hypothesis

$$H_o : P_T - P_R = 0 \text{ vs. } H_a : P_T - P_R \geq 24\%$$

A statistically significant difference indicates that the new drug is better than the control.

Effectiveness of Drug Products Decision Based on Results

True State	No difference	Minimal difference of 24%
No difference	Correct	Type Ⅰ Error (false positive)
Minimal difference of 24%	Type Ⅱ Error (false negative)	Correct

Effectiveness of Drug Products Decision Based on Results

- Significance level: The consumer's risk

  The chance that the decision based on the results there is a minimal difference of 24% improvement between t-PA and placebo when in fact there is no difference.

- Power = 1 – producer's risk

  The chance that decision based on the results concludes a minimal difference of 24% improvement between t-PA and placebo in fact there is.

Effectiveness of Drug Products Statistical Testing Procedures

- Step1

  State the null and alternative hypotheses

  - Null hypothesis: the one to be questioned

    No difference in the proportions of patients with improvement between t-PA and placebo.

  - Alternative hypothesis: the one of particular interest to investigators

    The minimal difference in the proportions of patients with improvement between t-PA and placebo is at least 24%.

Effectiveness of Drug Products Statistical Testing Procedures

- Step 2

  Choose an appropriate test statistics such as two-sample Z-statistic or t-statistic.

- Step 3
  - Select the nominal significance level

    the risk of type I error you are willing to commit

    Usually 5%

Effectiveness of Drug Products Statistical Testing Procedures

- Step 4
  - Determine the critical value, rejection region and decision rule

    For large samples, two-sided alternative and $\alpha = 0.05$, the critical value is z(0.025) = 1.96 and rejection region will be the one such that the absolute value of the test statistic is greater than 1.96.
  - Decision rule

    reject the null hypothesis if the resulting test statistic is in the rejection region.

Effectiveness of Drug Products Statistical Testing Procedures

Step 1 to step 4 should be determined and pre-specified in the Statistical Method section of the protocol before initiation of the study.

Effectiveness of Drug Products Statistical Testing Procedures

- Step 5

  When the study is completed, complete the value of the test statistic specific in Step 2 (protocol).

- Step 6

  Make decision based on the resulting value of the test statistic and decision rule specified in Step 4 (protocol).

Effectiveness of Drug Products Statistical Testing Procedures

- Conclusion
  - Reject the null hypothesis

    The sampling error is an unlikely explanation of discrepancy between the null hypothesis and observed values and the alternative hypothesis is proved at a risk of 5%.
  - Fail to reject null hypothesis

    The sampling error is a likely explanation and the data fail to provide sufficient evidence to doubt the validity of the null hypothesis.
  - Do **NOT** claim that the null hypothesis is accepted.

Effectiveness of Drug Products
P - value

- If there is no difference in in the proportions of patients with improvement between the two groups (i.e., the null hypothesis is true), the chance of obtaining a mean difference at least as large as the observed mean difference.
- If p-value is small, it implies that the observed difference is unlikely to occur if there is no difference in the proportions of patients with improvement between t-PA and placebo.

Effectiveness of Drug Products
P - value

- How small the p-value is sufficient enough to conclude that there exists a true difference in the proportions of patients with improvement between t-PA and placebo?
- It depends upon the risk that the investigator is willing to take for committing type I error.
- Nominal significance level = risk of type I error (The chance of concluding existence of a true difference in the proportions of patients with improvement between t-PA and placebo when in fact there is no difference)

Effectiveness of Drug Products
P - value

- If the observed p-value < the nominal significance level (i.e., the observed p-value < risk of type Ⅰ error), then conclude there exists a true difference in the proportions of patients with improvement between t-PA and placebo
- The nominal significance level = 5% or 1%
- The p-value for the observed difference in the proportions of patients with improvement between t-PA and placebo is 0.015.
- If the nominal significance level is 5%, then it is concluded that there is a difference in the proportions of patients with improvement between t-PA and placebo in target population of patients with acute ischemic stroke .

Effectiveness of Drug Products
P - value

- We can not make the same decision if the nominal significance level is chosen to be 1%.
- Should always reported the observed p-value and let readers and reviewers judge the strength of evidence by themselves and do not use p-value < 0.05.

Equivalence of Generic Drug Products

- New Drug Development (Innovative Drugs)
- Length: an average of 12 years
- Cost: an average of 800 million US dollars
- Success rate:
  - 1 out of 10000 molecules screened
  - 60% failure rate during clinical development
  - 30% success rate for phase III trials

Equivalence of Generic Drug Products

- Abbreviated New Drug Application (Generic Drugs)
- After the patent of the innovative drug is expired, all other manufacturers can produce the same drug product
- Patents of most innovative drugs expires by 2005: big market
- Requires evidence of bioequivalence between innovative and generic drug products

Equivalence of Generic Drug Products

- Pharmacokinetic Measures
  - Absorption
  - Distribution
  - Metabolism
  - Elimination

  Based on the plasma concentrations of active ingredients C_0, C_1,…,C_K measures at 0, t_1,…,t_k.

Equivalence of Generic Drug Products

- Total Exposures
  - AUC (0-t_K), AUC (0-∞)
- Peak Exposure
  - C_{max} – peak drug concentration
- Partial Exposure
  - Partial AUC: AUC(0-t_i)
- Other Measures
  - Half-life ($t_{1/2}$), % fluctuation.

Equivalence of Generic Drug Products

Equivalence hypothesis

$\theta = \mu_T - \mu_R$

$H_o: \mu_T - \mu_R \leq \theta_L$ or $\mu_T - \mu_R \geq \theta_L$

vs. $H_a: \theta_L < \mu_T - \mu_R < \theta_U$

Equivalence of Generic Drug Products - Average Bioequivalence

Two one-sided hypotheses:

$$H_{oL}: \mu_T - \mu_R \leq \theta_L \text{ vs. } H_{aL}: \mu_T - \mu_R > \theta_L$$

and

$$H_{oU}: \mu_T - \mu_R \geq \theta_U \text{ vs. } H_{aU}: \mu_T - \mu_R < \theta_U$$

The parameter space of H_o is the union of the parameter spaces of H_{oL} and H_{oU}.

The parameter space of H_a is the intersection of the parameter spaces of H_{aL} and H_{aU}.

Equivalence of Generic Drug Products - Average Bioequivalence

Schuirmann's Two One-sided Tests Procedure (TOST, 1987)

Conclude ABE if

$$T_L = (f - \theta_L)/v(f) > t(\alpha, n_1+n_2-2)$$

and

$$T_U = (f - \theta_U)/v(f) < -t(\alpha, n_1+n_2-2),$$

where f is the LSE for θ

Equivalence of Generic Drug Products - Average Bioequivalence

Confidence Interval Approach

If a $(1-2\alpha)100\%$ confidence interval for the difference $\mu_T - \mu_R$ or the ratio μ'_T/μ'_R is within the acceptance limits as recommended by the regulatory agency, then accept the test formulation; otherwise reject it. Westlake (1981)

$\alpha = 5\% \Rightarrow 90\%$ C.I.

log-scale: $\mu_T - \mu_R$: ± 0.2231

Original Scale: μ'_T/μ'_R: (80%, 125%)

TOST is operationally equivalent to CI approach

This is the requirement by most of health regulatory agencies in the word

Estimation of Shelf-life

- Shelf-life (expiration dating period)

  Time interval during which a drug product is expected to remain within the specifications, provided that it is stored under the conditions defined on the container label

- Expiration date

  The date placed on the container label of a drug product designating the time prior to which a batch of the product is expected to remain within the approved shelf life, if stored under defined conditions, and after which it must not be used.

Estimation of Shelf-life

ICH Q1A(R2) guidance (2003) P.16

"An approach for analyzing data of quantitative attribute that is expected to change with time is to determine the time at which the 95% one-sided confidence limit for the mean curve intersects the acceptance criterion"

ICH Q1E guidance (2004) p.11

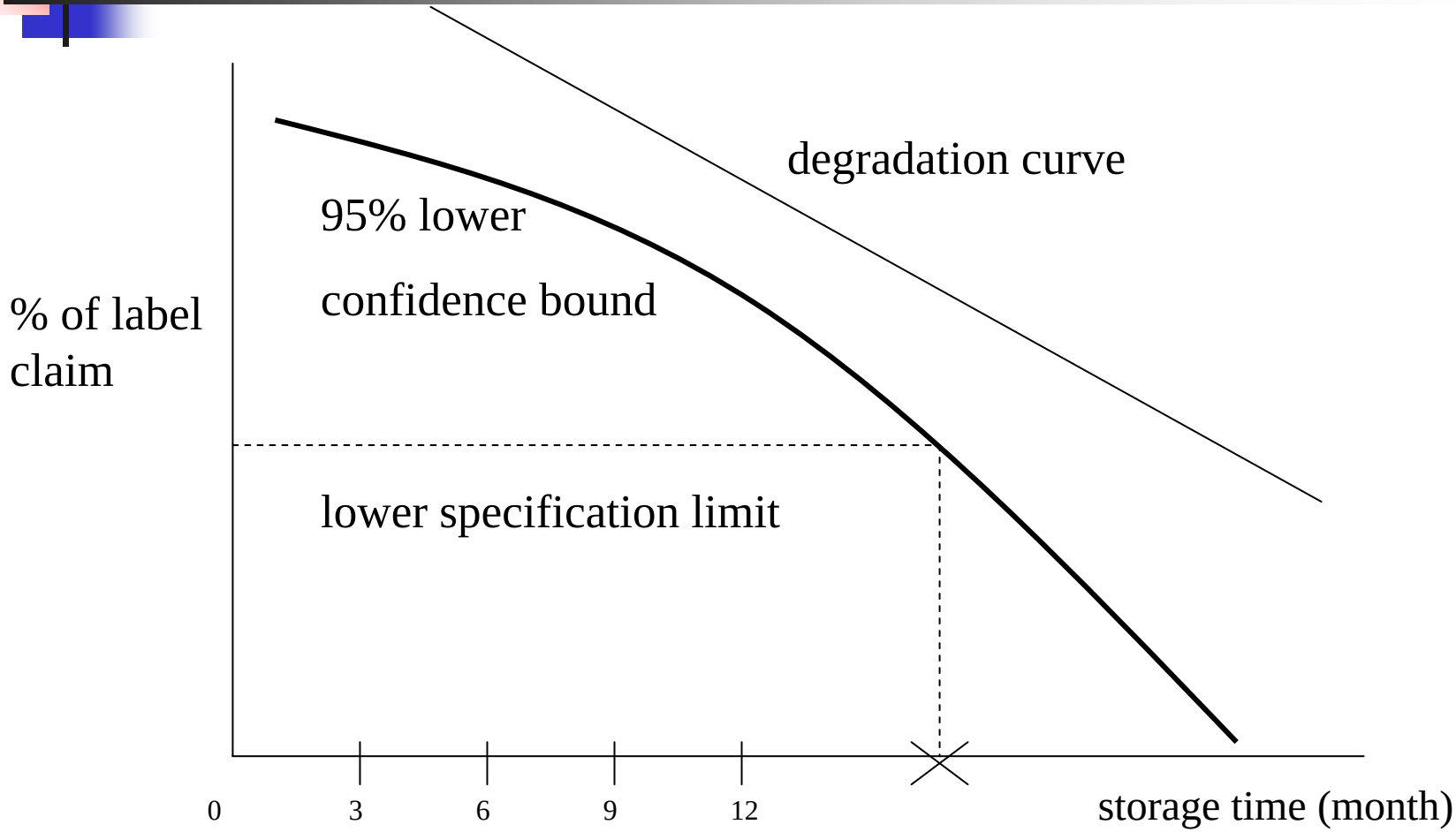
A two-sided 95% confidence interval or 95% one-sided upper or lower confidence interval can be also used.

One-sided lower limit: known degradation

One-sided upper limit: known impurities

Two-sided interval: unknown situation about increase or decrease of the assay with the time

Estimation of Shelf-life



Estimation of Shelf-life

- Only consider the case where the drug product characteristic decreases linearly with time.

- **Model:** $Y_j = \alpha + \beta X_j + \varepsilon_j,\ j = 1,2,...,n$

Y_j : jth response of assay at time X_j,

α : Intercept(batch effect),

β : Slope(degradation rate),

X_j: time at which Y_j is observed,

ε_j : random error ~ $N(0,\sigma^2)$.

Estimation of Shelf-life

Construct (l-2α)100% C.I. for X for which the pth upper quantile of the distribution of Y given X is equal to some specified value η .

The pth upper quantile of the distribution of Y given X is $\alpha + \beta X + \sigma z_p$, where z is the pth upper quantile of a standard normal distribution.

The value of X for which the hypothesis

$$H_0: [(\eta - \alpha - z_p \sigma)/ \beta] \leq X$$

is not rejected at the 2α significance level will constitute an(l-2α)100% C.I. for X.

Esterling(1969)

Estimation of Shelf-life

- Stability study: mean degradation => p=0.5 => z_p=0.

H_0: $[(\eta - \alpha)/\beta] \leq X$

=> H_0: $\eta - \alpha - \beta X \leq 0$

H_a: $\eta - \alpha - \beta X > 0$

=> H_0: $\alpha + \beta X \geq \eta$

H_a: $\alpha + \beta X < \eta$

=> H_0: $(\eta - \alpha)/\beta \leq X$

H_a: $(\eta - \alpha)/\beta > X$

Estimation of Shelf-life

- Stability study: mean degradation => p=0.5 => z_p=0.

  H_0: $[(\eta - \alpha)/\beta] \leq X$

  => H_0: $\eta - \alpha - \beta X \leq 0$

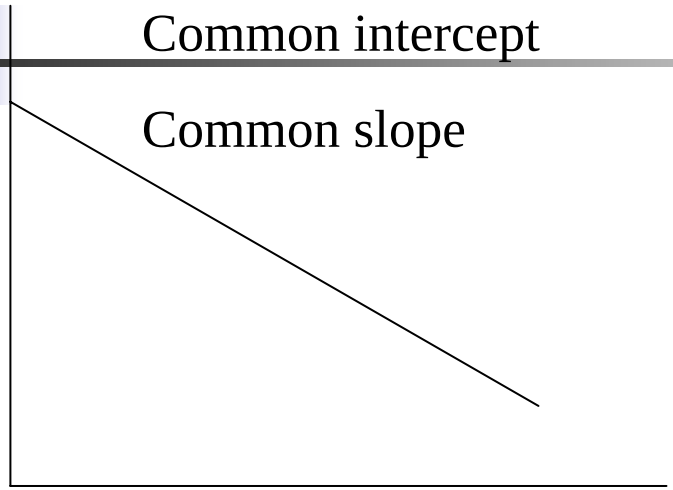
The set of values of X for which H_0 is not rejected at the 2α significance level is

$A = \{X: [\eta - (a + bX)]^2 \leq t^2_{(\alpha, n-2)}[SE(X)]^2\}$

Estimation of Shelf-life

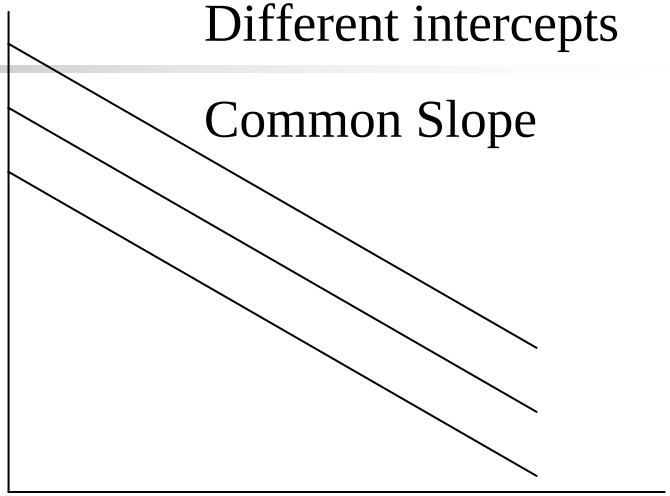
Common intercept

Common slope



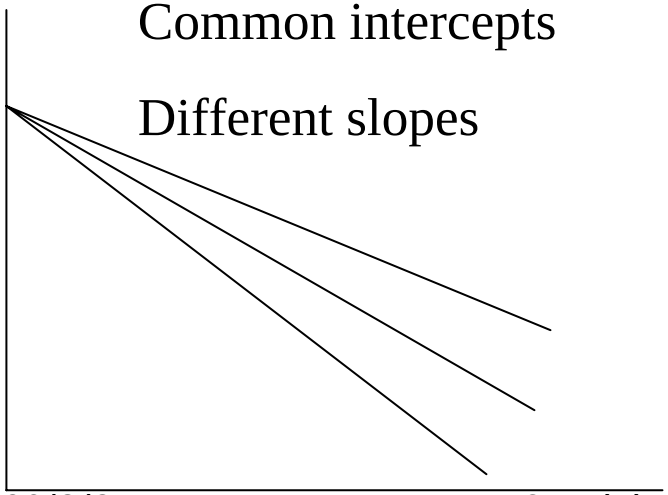
Different intercepts

Common Slope



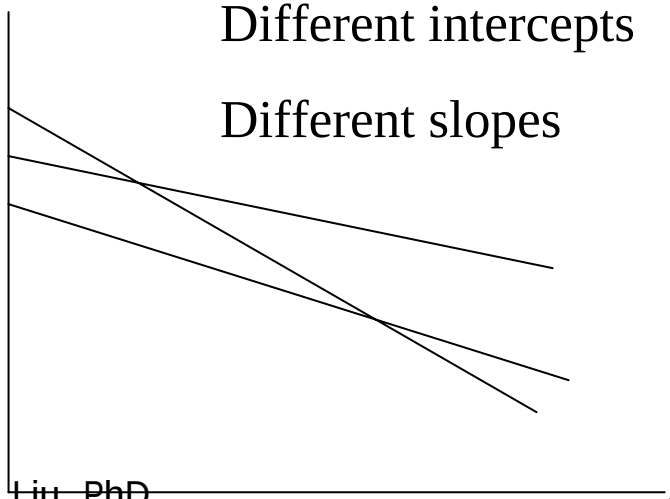
Common intercepts

Different slopes



Different intercepts

Different slopes



Quality Control of Drug Products

- Sampling Plan and Acceptance Criteria
  - Content uniformity of dosage units
    - USP/NF general chapter[905]
  - Dissolution Testing
    - USP/NF general chapter[711]
  - Disintegration Testing
    - USP/NF general chapter[701]

Disintegration Testing
USP/NF general chapter [711]

TABLE 5.2.3 Acceptance Criteria for Disintegration

Stage	Number tested	Pass if:
S_1	6	All of the units have disintegrated completely.
S_2	12	No less than 16 of the total of 18 units ($S_1 + S_2$) tested disintegrate completely.

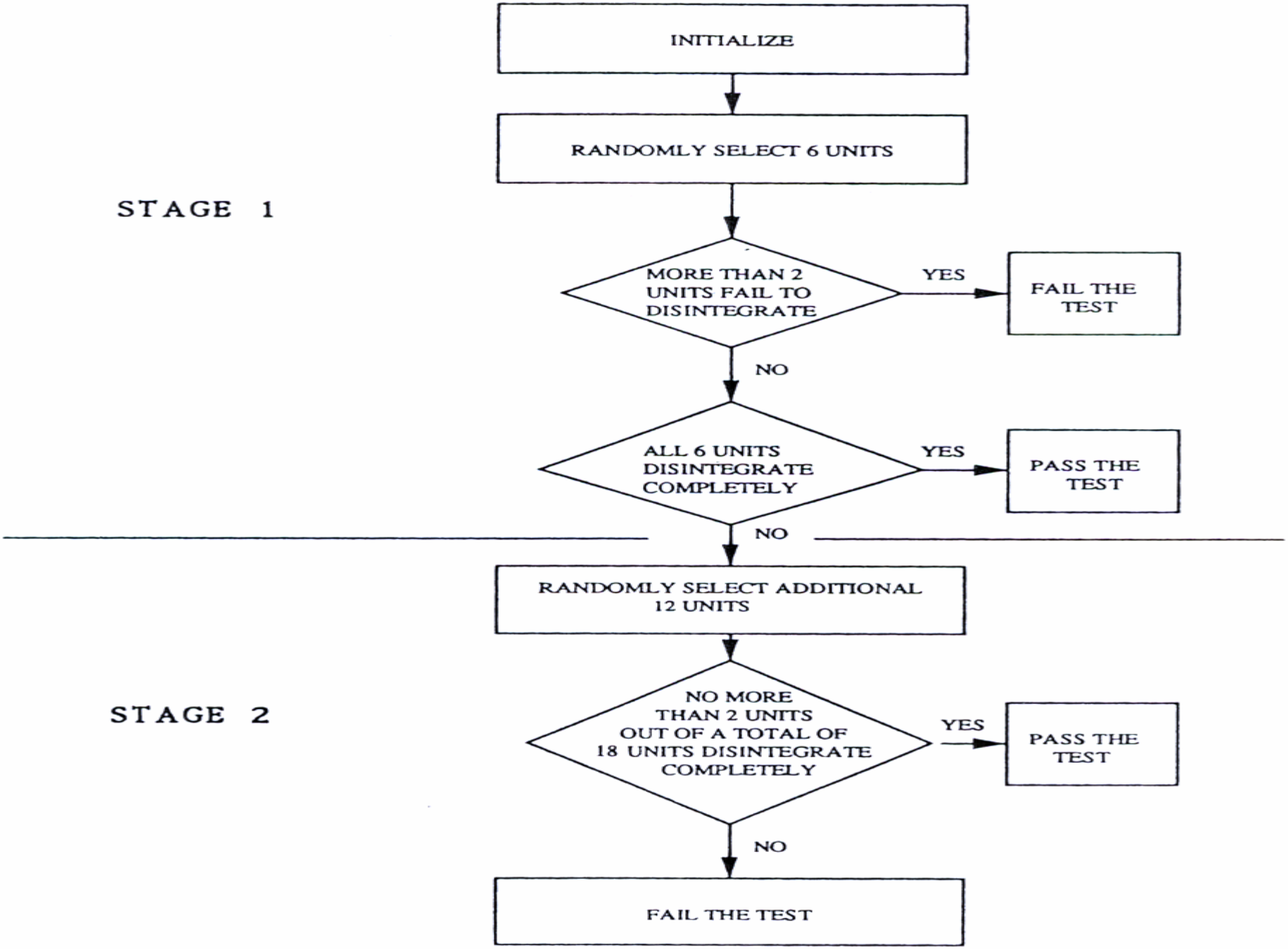


FIGURE 5.2.3 Flowchart of test procedure for disintegration.

Disintegration Testing

Let Y be the disintegration time. Again we assume that Y follows a normal distribution with mean μ and variance σ^2 .

- Also, let $p = P\{0 < Y <$ UL$\}$,
  where UL denotes the specified limit. Since the disintegration test involves only one acceptance criterion at both stages of the sampling plan, the exact probability can be computed. Let

  C_{11} = {all six units disintegrate completely},

  C_{12}= {one unit fails to disintegrate completely},

  C_{13}= {two units fail to disintegrate completely},

  C_{21} = {11 of 12 additional units disintegrate completely},

  C_{22} = {all 12 additional units disintegrate completely}.

Then the exact probability of passing the disintegration test is given as follows:

$$\begin{aligned}
\boldsymbol{P}\{\text{pass}\} &= \boldsymbol{P}\{\mathrm{C}_{11}\} + \boldsymbol{P}\{\mathrm{C}_{21} + \mathrm{C}_{22} \mid \mathrm{C}_{12}\}\boldsymbol{P}\{\mathrm{C}_{12}\} \\
&\quad + \boldsymbol{P}\{\mathrm{C}_{22} \mid \mathrm{C}_{13}\}\boldsymbol{P}\{\mathrm{C}_{13}\} \\
&= \boldsymbol{p}^6 + \left\{\binom{12}{11}\boldsymbol{p}^{11}(1-\boldsymbol{p}) + \boldsymbol{p}^{12}\right\}\left\{\binom{6}{1}\boldsymbol{p}^5(1-\boldsymbol{p})\right\} \\
&\quad + \mathrm{p}^{12}\left\{\binom{6}{2}\boldsymbol{p}^4(1-\boldsymbol{p})^2\right\} \\
&= \boldsymbol{p}^6 + 6\boldsymbol{p}^{17}(1-\boldsymbol{p}) + 87\boldsymbol{p}^{16}(1-\boldsymbol{p})^2.
\end{aligned}$$

It can easily be verified that if the desired probability of passing the disintegration test is 0.5, p is approximately about 0.831. If, in addition, the specified time limit, UL, is 30 min, it follows that

$$\begin{aligned} \boldsymbol{p} &= \boldsymbol{P}\{\boldsymbol{Y} < \boldsymbol{UL}\} = \boldsymbol{P}\{\boldsymbol{Y} < 30\} \\ &= \boldsymbol{P}\{\frac{\boldsymbol{Y}-\mu}{\sigma} < \frac{30-\mu}{\sigma}\} = \boldsymbol{P}\{\boldsymbol{Z} < \boldsymbol{Z}(0.169)\} \\ &= 0.831 \end{aligned}$$

where Z is a standard normal variable and Z(0.169) is the 16.9% upper quantile of a standard normal distribution. Therefore

$$\frac{30-\mu}{\sigma} = \boldsymbol{Z}(0.169) = 0.957$$

Hence the contour for μ and σ^2 is a linear decreasing function of given by $0.957\sigma = 30 - \mu$

where 0.957=Z(0.169)

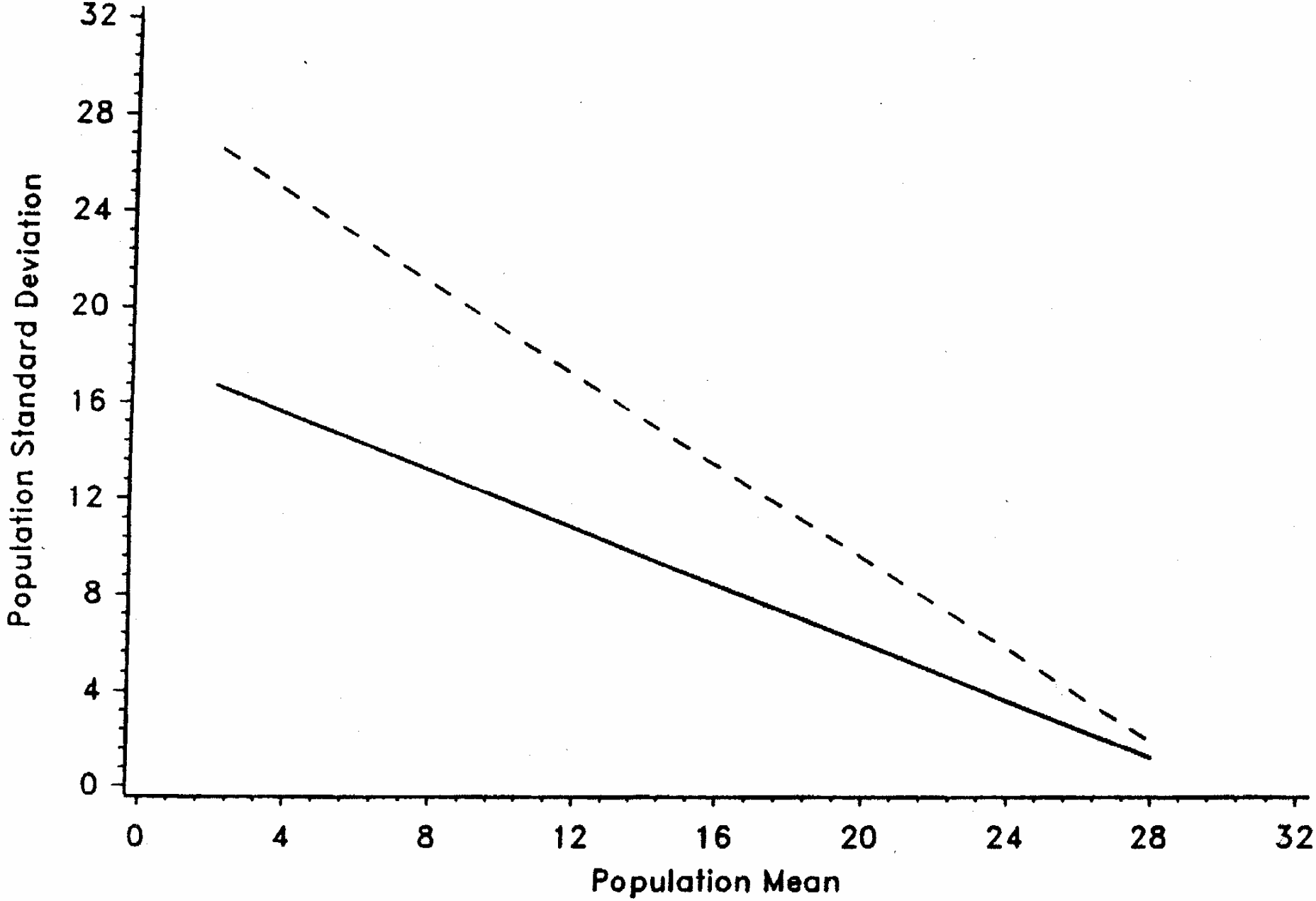
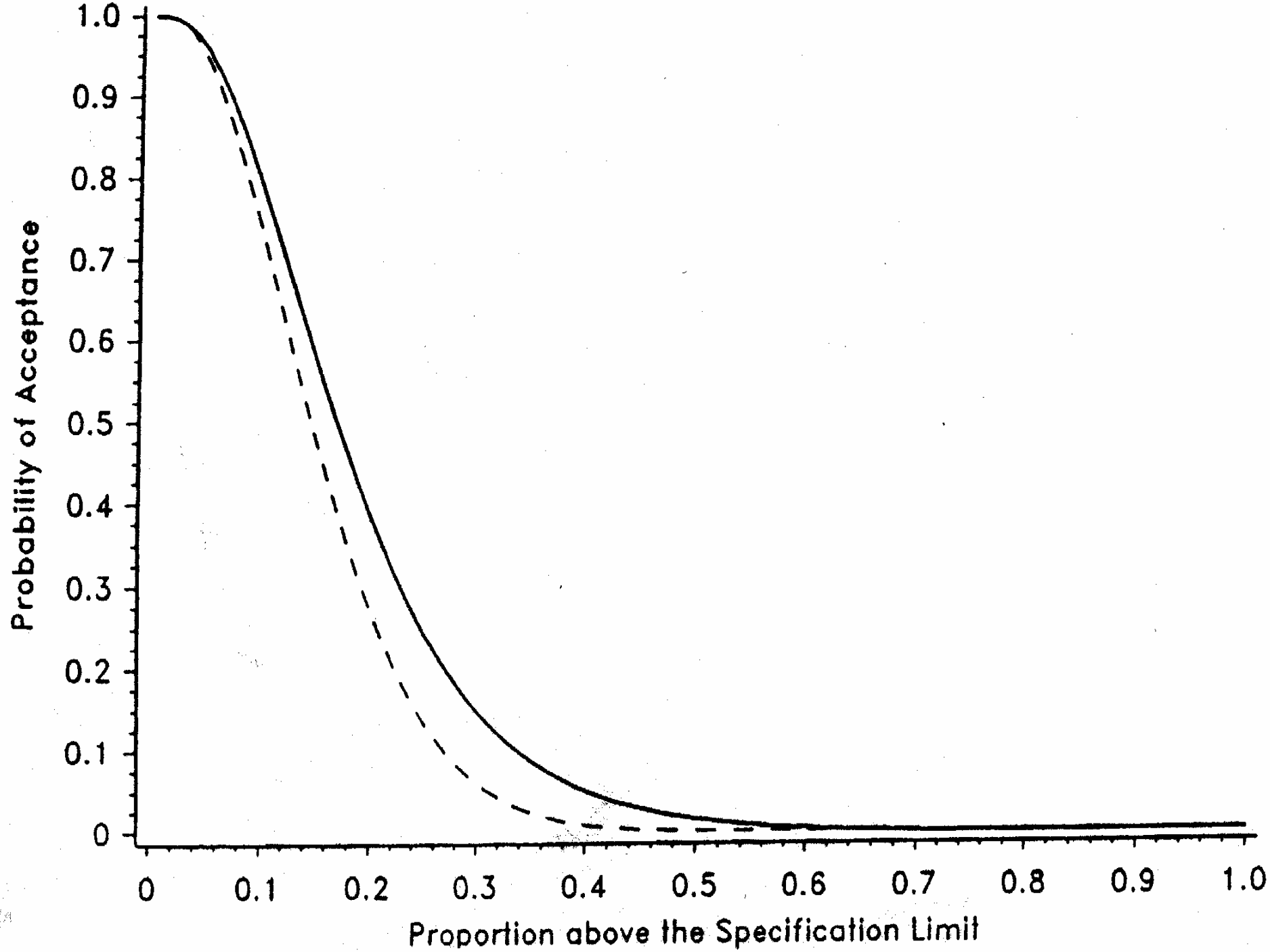


FIGURE 5.3.4 Lower probability bounds for passing disintegration test. Solid curve, lower probability = 95%; dashed curve, lower probability = 50%.


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Probability of Acceptance
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Proportion above the Specification Limit

Evaluation of Diagnostic Devices

Simplest Situation: Binary Outcomes from marker test (+, -)
Binary Classification of Disease (Yes, No)

Design Matrix for Diagnostic Marker Tests

	Correct ("Gold Standard") True State of Disease		
Diagnosis Made from Marker Test	Present (D)	Absent (D)	Total
Positive (T)	a (1-β)	b (α)	m_1
Negative (T)	c (β)	d (1-α)	m_2
Total	n_1	n_2	N

Evaluation of Diagnostic Devices

Retrospective Sampling Plan (case-control)

- **Sensitivity** (True Positive rate): Capacity for making a correct diagnosis in subjects with the disease
- Estimated Sensitivity:

  100% x a/(a+c)
- **Specificity** (True Negative rate): Capacity for making a correct diagnosis in subjects without disease
- Estimated Specificity:

  100% x d/(b+d)

Evaluation of Diagnostic Devices

- **Positive Predictive Value** (Positive Predictive Accuracy): the proportion of subjects with the disease given the positive results.
  = 100% x a/(a+b)
- **Negative Predictive Value** (Negative Predictive Accuracy): the proportion of subjects without the disease given the negative results.
  = 100% x d/(c+d)
- **False positive rate**: given the positive results ,the proportion of subjects without the disease
  =1 – positive predictive value = 100% x b/(a+b)
- **False negative rate**: given the negative results, the proportion of subjects with the disease
  = 1 – negative predictive value = 100% x c/(c+d)

Evaluation of Diagnostic Devices

Other Definitions of False Positive Rate and False Negative Rate

False positive rate : given the subjects without thedisease, the proportion of subjects with positive results = $b/(b+d) = b/n_2$

False negative rate : given the subjects with the disease, the proportion of subjects with negative results = $c/(a+c) = c/n_1$

False positive rate = 1 - specificity

False negative rate = 1 - sensitivity

Evaluation of Diagnostic Devices

Example (Feinstein, 2002)

New Maker Test Result	Diseased Cases	Nondiseased Control	Total
Positive	46	2	48
Negative	4	48	52
Total	50	50	100

Evaluation of Diagnostic Devices

Data from Example 2 (Feinstein, 2002)

- Sensitivity = 100% x 46/50 = 92.0%
- Specificity = 100% x 48/50 = 96.0%
- Prevalence = 100% x 50/100 = 50.0%
- Positive Predictive Value
  = 100% x 46/48 = 95.8%
- Negative Predictive Value
  = 100% x 48/52 = 92.3%
- False Positive Rate = 100% x 2/48 = 4.2%
- False Negative Rate = 100% x 4/52 = 7.7%

Evaluation of Diagnostic Devices

Type of Diagnostic Markers

- Binary Test Results (+,-)
- Multiple Categorical Results
  - Abnormality Rating
  - Severity Rating
  - Urine test: None, trace, 1+, 2+
  - HER2 test: 0, 1+, 2+, 3+
- Continuous Test Results
  - PSA
  - Intraocular Pressure
  - Glucose tolerance test
  - Gene expression level

Evaluation of Diagnostic Devices

- To convert a ranking scale or a continuous measurement into a binary outcomes (+,–), we need a cutoff point or threshold.
- ***Example:***
- FBG > 126mg/dL DM (+)
- ≤ 126mg/dL DM (–)
- S-T Depression in Exercise Stress Test
- Class D < 1.5 min CAD (+)
- ≥ 1.5 min CAD (–)

Evaluation of Diagnostic Devices

At a specific threshold, relationship of sensitivity, specificity, false positive and false negative rates can be interpreted through hypothesis testing:

H0:Absence of the disease

H1:Presence of the disease

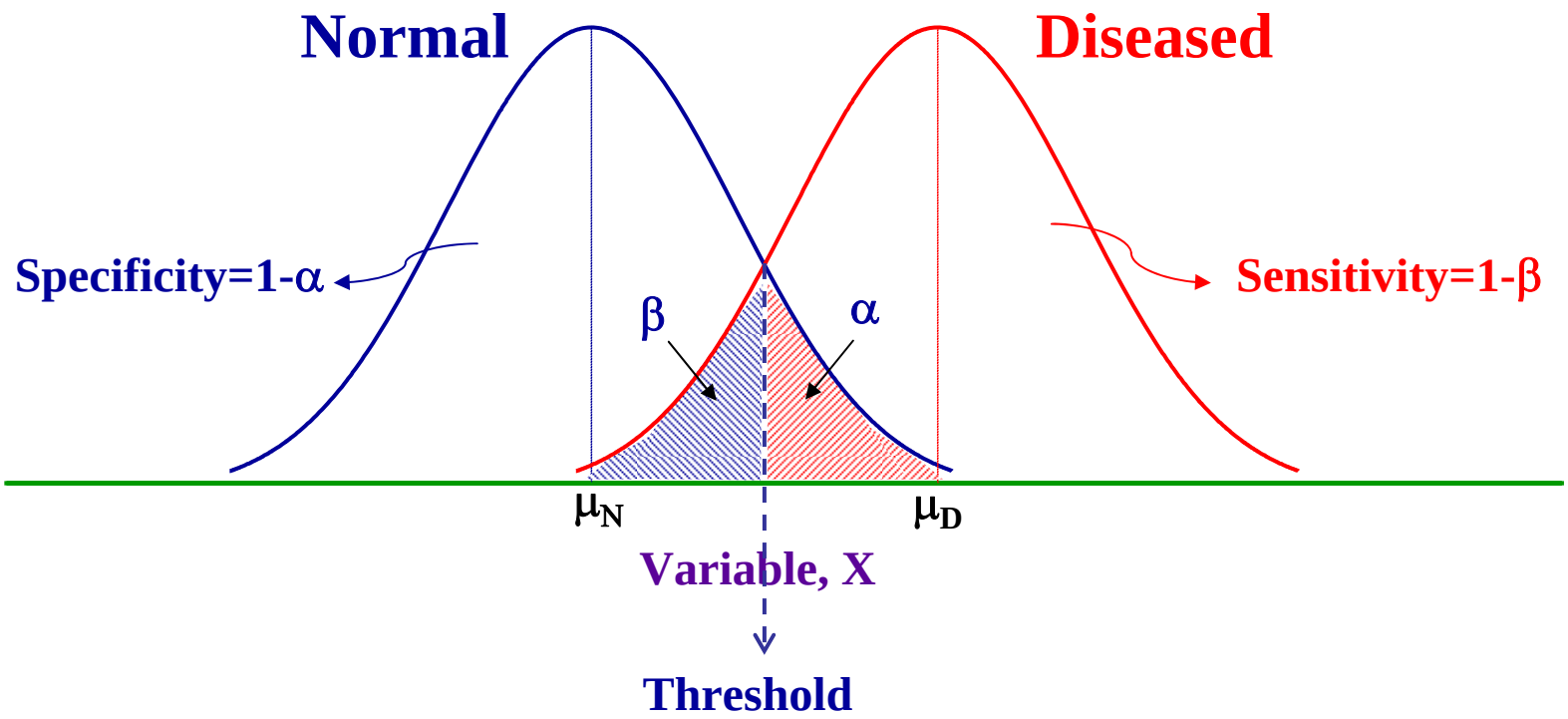
α =Pr[Type I Error]

=Pr[test positive | no disease]

β=Pr[Type II Error]

=Pr[test negative | disease]

Evaluation of Diagnostic Devices



Evaluation of Diagnostic Devices

Sensitivity = Pr[test positive | disease]
= $1 - \beta$
= power of the statistical procedure

Specificity = Pr[test negative | no disease]
= $1 - \alpha$

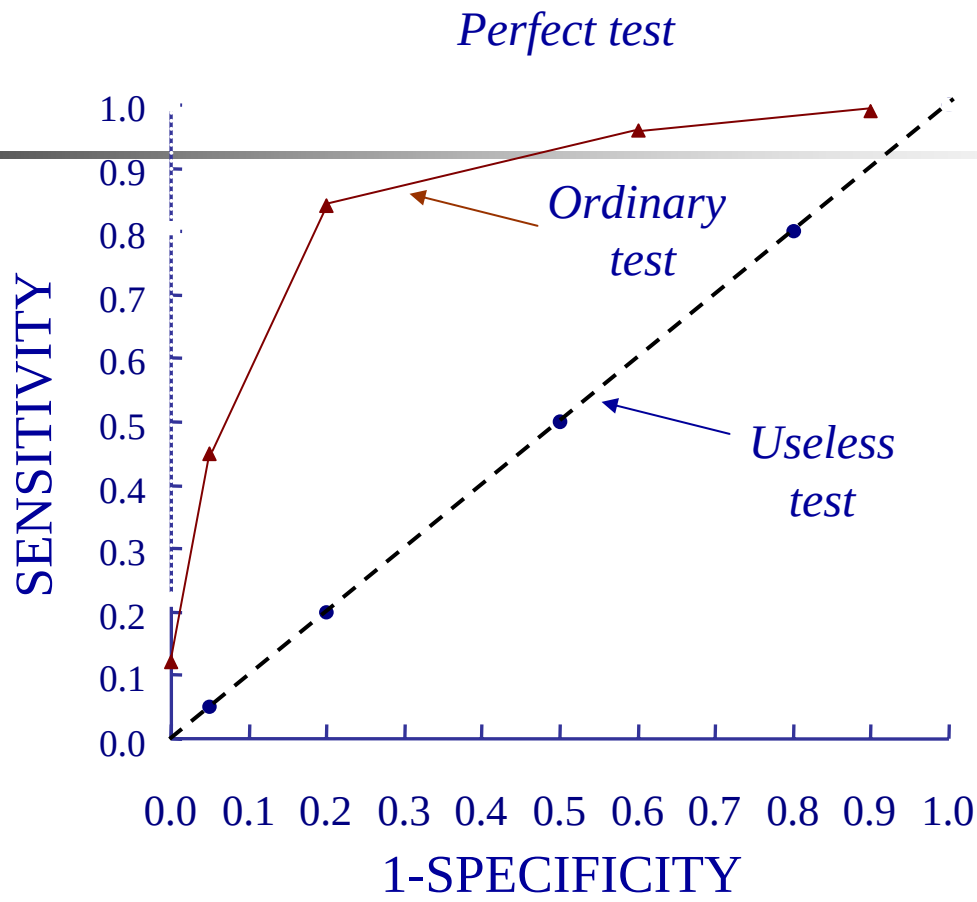
- $\alpha\uparrow \Rightarrow \beta\downarrow \Rightarrow (1-\beta)\uparrow$
- A test with a high sensitivity also has a high incorrect positive rate but a low incorrect negative rate. A test with a high specificity also has a high incorrect negative rate but a low incorrect positive rate.

Evaluation of Diagnostic Devices

- At each individual threshold (cut-off), sensitivity and specificity can be computed.
- A Receiver Operating Characteristic (ROC) curve is a graphic presentation of sensitivity against 1-specificity.
- It is a path in the unit square, from the lower left corner to the upper right corner. In fact, it can be viewed as a cumulative distribution function.
- Swets (1979), Hanley and McNeil (1982), Metz (1978, 1980)

Evaluation of Diagnostic Devices

- In a useless marker test, the ROC curve will be a straight line at a 45^{o} angle.
- The area under the ROC curve provides a summary index for diagnostic accuracy across over all possible values of thresholds.
- The range of the area under the ROC curve is from 0.5 (50%) to 1.0(100%)
- In a useless marker test, the area under the ROC curve is 50% which is the same as flopping a fair coin.
- For non-inferiority or equivalence test based on the paired ROC curve area, see Liu, et al. (2005, *Statistics in Medicine*)



Source: Feinstein (2002)

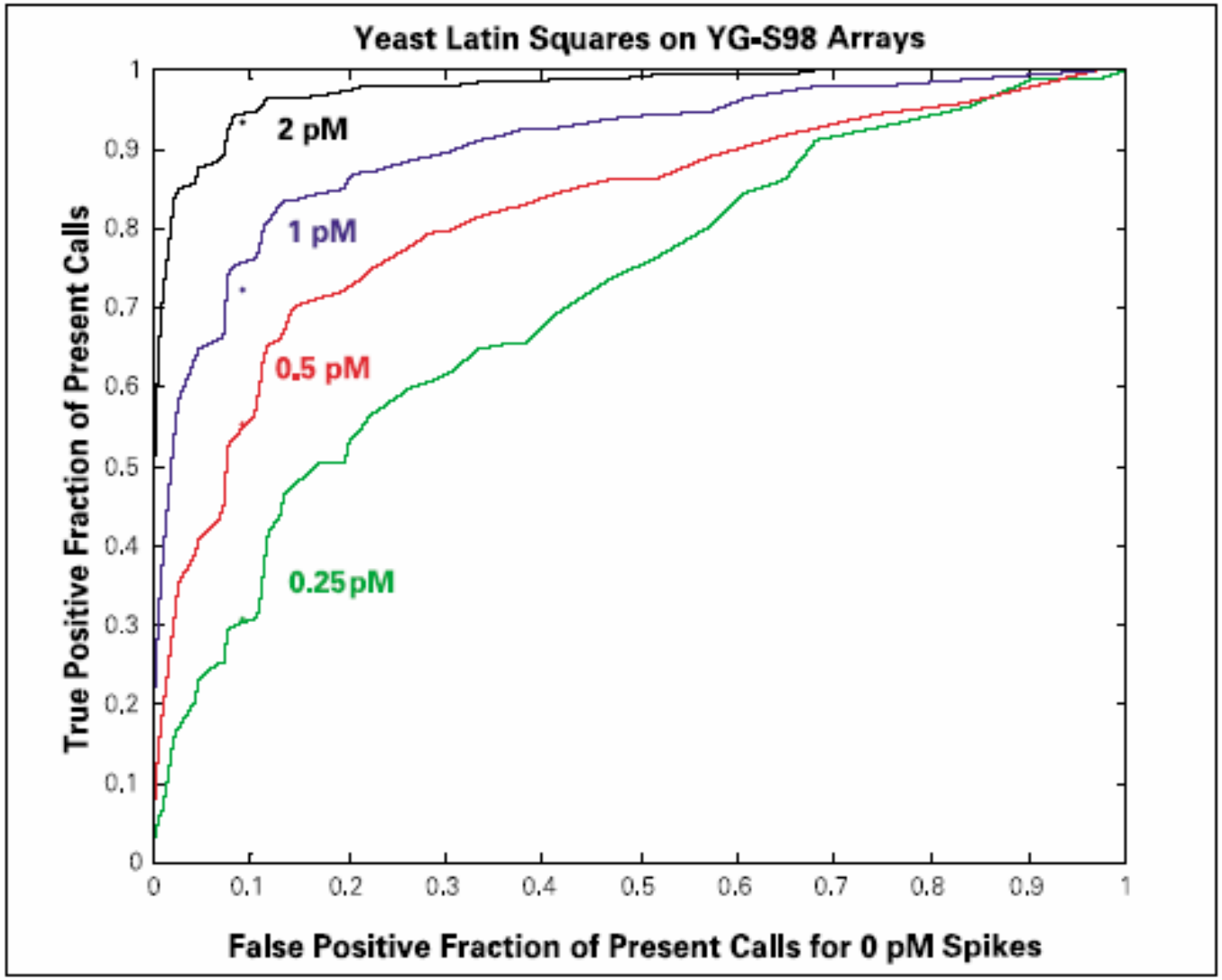
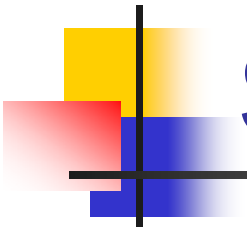


Figure 1. ***Detection Calls in Yeast Spikes: balancing sensitivity and specificity.***

From Affymetrix Technical Note 2 “Fine tuning your data analysis”



Summary

- Descriptive Statistics
  - Description of characteristics and estimation of special attributes of drug and device products
- Inferential Statistics
  - Decision-making tool for approval of drug and device products for marketing

References

- Chow, SC and Liu, JP (2004) Design and Analysis of Clinical Trials, 2nd Ed. Wiley
- Chow, SC and Liu, JP (2000) Design and Analysis of Bioavailability and Bioequivalence Studies, Marcel Dekker, Inc.
- Chow, SC and Liu, JP (1995) Statistical Design and Analysis in Pharmaceutical Sciences, Marcel Dekker, Inc.