

---

# Introduction to Statistics for Pathologists with Emphasis on Evaluating Lab Tests

Robin T. Vollmer, MD MS

# The Problems

---

- ◆ What is the value of serum PSA in prostate cancer? How should one use PSA? Does PSA work the same for blacks and whites?
- ◆ How good is urine cytopathology as lab test for Ca?
- ◆ FNA of Thyroid Lesions: which microscopic details are most important?

# The Problems

---

- ◆ CK-MB vs Troponin I for dx of AMI
- ◆ Serum fragments of cytokeratin 19 as a measure of tumor burden in non-small cell lung cancer. (CYFRA 21-1)
- ◆ Importance of number of lymph nodes examined in N0 colon cancer.

# The Basic Concepts

---

- ◆ Probability
- ◆ Random Variables
- ◆ Statistical Tests
- ◆ Multivariate Models

# Probability: The Beginning

---

- ◆ Events,  $E$ : Outcomes of an experiment or set of observations
- ◆ Probability,  $P$ :
- ◆ A mapping of  $E \rightarrow [0,1]$

## Relative frequency concept of probability:

---



no. with event E



$P(E) \sim$  -----



total no. observed

# Disease and Test Events

---

- ◆ Sample Space: set of patients
- ◆ Event: presence of disease,  $D^+$
- ◆ Event: presence of pos. test,  $T^+$
- ◆  $P(D^+)$ : probability of disease
- ◆  $P(T^+)$ : probability of pos. test

# Probability of 2 Events, D+ & T+

---

- ◆ no. with D+ and T+
- ◆  $P(D+ \& T+) \sim \frac{\text{no. with D+ and T+}}{\text{total no.}}$
- ◆ total no.
- ◆  $P(D+ \text{ or } T+) = P(D+) + P(T+) - P(D+ \& T+)$
- ◆

# Mutually Exclusive Events

---

- ◆ If  $D^+$  and  $T^+$  are mutually exclusive, then
- ◆  $P(D^+ \& T^+) = 0$ .
- ◆ In general  $D^+$  and  $D^0$  are mutually exclusive, and
- ◆  $P(D^0) = 1 - P(D^+)$

# Conditional Probabilities

---

◆  $P(D+|T+)$  reads the probability of observing event  $D+$ , given that event  $T+$  has occurred.

◆  $P(D+ \& T+)$

◆  $P(D+|T+) = \frac{P(D+ \& T+)}{P(T+)}$

◆  $P(T+)$

◆  $\sim (\text{no. with } D+ \& T+) / (\text{no. with } T+)$

# Conditional Probability

---

$$P(T+|D+) = \frac{P(D+ \& T+)}{P(D+)}$$

- ◆  $\sim$  (no. with D+ & T+) / (no. with D+)

## In other words:

---

- ◆  $P(D+ \& T+) = P(D+|T+) * P(T+)$

- ◆ Which also equals:

- ◆  $= P(T+|D+) * P(D+)$

- ◆ (This comes from the commutative nature of probability as a set function)

# Independent Events

---

- ◆ 2 Events D+ and T+ are statistically independent if and only if:
- ◆  $P(D+ \& T+) = P(D+) * P(T+)$
- ◆ i.e.  $P(D+|T+) = P(D+)$
- ◆ And  $P(T+|D+) = P(T+)$

# Random Variables

---

- ◆ Random variables come into play when the events are numerical or can be represented by numbers.
- ◆ e.g. age, weight, height

# Spread Sheet Example

---

◆ Pt	x1	x2	x3	x4	x5	x6
◆ 1	1	5.0	1	10.	25	2
◆ 2	1	4.5	1	9.5	30	1
◆ 3	0	6	1	9	45	4

# Categories of Random Variables

---

- ◆ Binary: e.g. D+ and D0
- ◆ Ordinal: e.g. No. of recurrent bladder tumors before an invasive tumor event.
- ◆ Continuous: Time to failure, Serum PSA.
- ◆ The type of statistical analysis needed largely depends on the type of random variables present.

# Calculus of Probability and Random Variables

---

- ◆ The distribution function  $F(x)$  is the representation of Probability for  $x$ .
- ◆  $P(x \leq a)$  is the definition of  $F(a)$
- ◆ The probability density function,  $f(x)$ , is related to  $F(x)$  through calculus.

# Probability Density, $f(x)$

---

◆  $f(x)$  is a differential in probability:

◆  $P(a < x \leq a + dx)$

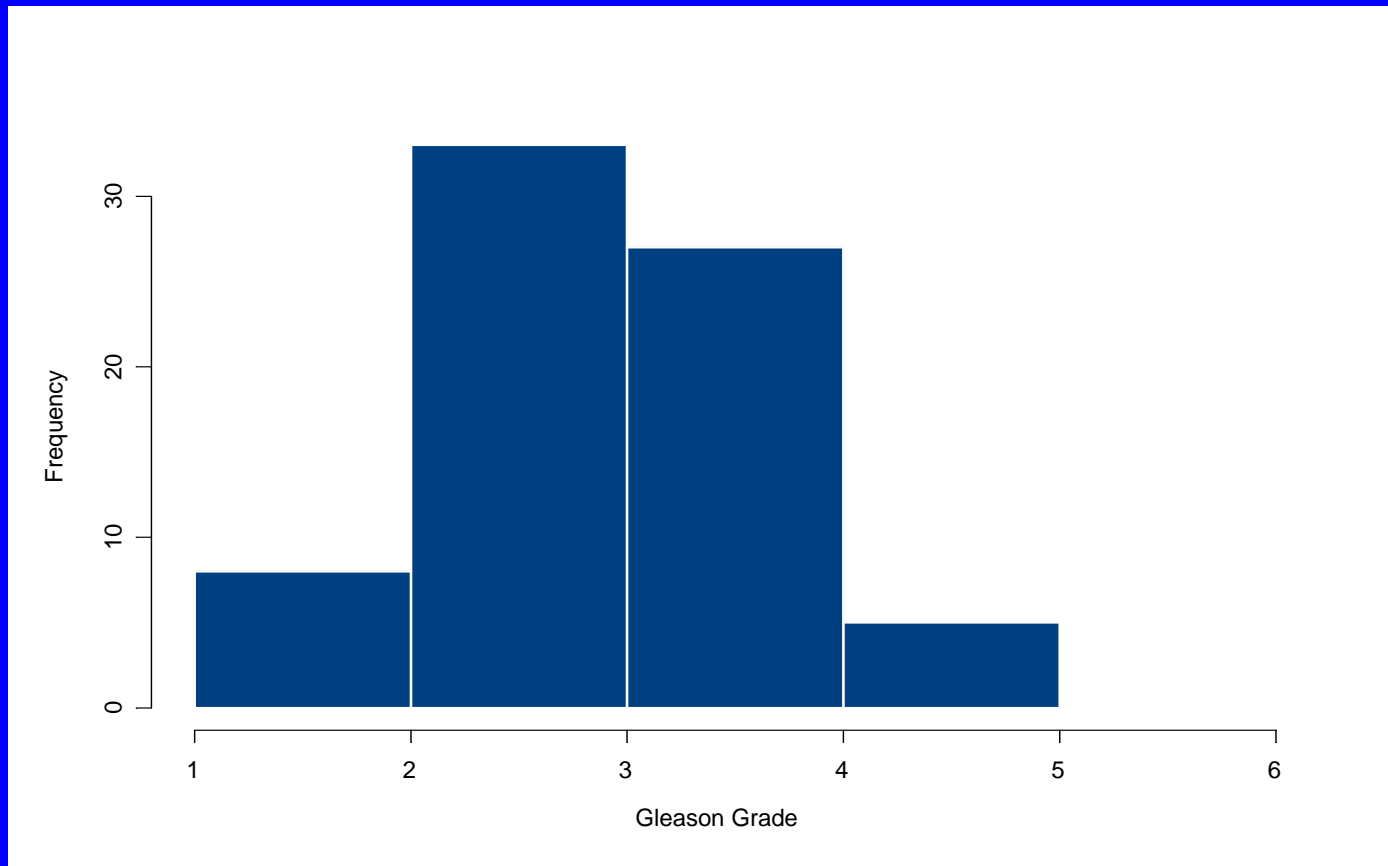
◆  $f(a) = \lim_{dx \rightarrow 0} \frac{\text{-----}}{dx}$

◆  $dx \rightarrow 0$   $dx$

◆ Useful to think of  $f(x)$  as a histogram.

# Histogram: A Graphical Illustration of $f(x)$

---



# Probability Densities, $f(x)$

---

- ◆ For many types of  $x$ ,  $f(x)$  are helpful mathematical functions
- ◆ Examples:
- ◆ binomial for binary  $x$ , i.e.  $\{0,1\}$
- ◆ normal, chi-square, exponential,  $F$ ,  $t$ , gamma, ..., for continuous  $x$ 's

# A Random Sample

---

- ◆ Suppose we observe a single random variable like presence and absence of ca (i.e. 1 or 0) on a series of pts.
- ◆ Thus, what we observe will be a series of 1's and 0's: 0,1,0,1,1,0, ...

# Spread Sheet Example

---

◆ Pt	Ca	x2	x3	x4	x5	x6
◆ 1	1	5.0	1	10.	25	2
◆ 2	1	4.5	1	9.5	30	1
◆ 3	0	6	1	9	45	4

# A Random Sample

---

- ◆ To be a random sample, the pts and their 0,1 outcomes must be statistically independent of one another.
- ◆ In other words:
- ◆  $P(\{\text{Ca in pt 1}\} \text{ and } \{\text{Ca in pt 2}\})$  must equal:
- ◆  $P(\text{Ca} \mid \text{pt 1}) \times P(\text{Ca} \mid \text{pt 2})$

# Statistic, s

---

- ◆ s: a numerical summary of the of observed sequence of random variables, one x for each pt., in a random sample
- ◆ Sample Mean =  $(\sum x_i) / n$
- ◆
- ◆  $\sum (x_i - \text{Sample mean})^2$
- ◆ Sample Variance =  $\frac{\sum (x_i - \text{Sample mean})^2}{n - 1}$
- ◆

# Other Statistics

---

- ◆ Mode is the most common value of  $x$ .
- ◆ Median is the value of  $x$  that divides the sample space into 2 equal halves.
- ◆ The Chi-square statistic: the sum of
- ◆  $[(\text{observed} - \text{expected}) / \text{expected}]^2$

# Serum PSA--the univariate approach

---

- ◆ What is the normal serum PSA?
- ◆ (what is the probability distribution function,  $F$ , for serum PSA?)

# Orthodox Reference Question:

---

- ◆ What is the range of values expected for PSA in 95% of the reference population?
- ◆ We seek limits  $a$  and  $b$  such that 95% of the ref. population fits between  $a$  and  $b$ .
- ◆ i.e. we seek values  $a$  and  $b$  such that  $F(a) = 0.025$  and  $(1-F(b)) = 0.025$

# Reference Ranges: Parametric

---

- ◆ Examine plot of frequency distribution.
- ◆ Normal ?: mean  $\pm$  2 standard deviations of  $x$ .
- ◆ Log-normal ?: mean  $\pm$  2 standard deviations of  $\log(x)$
- ◆ Exponential ?: 0 to  $3.0 * \text{mean}$

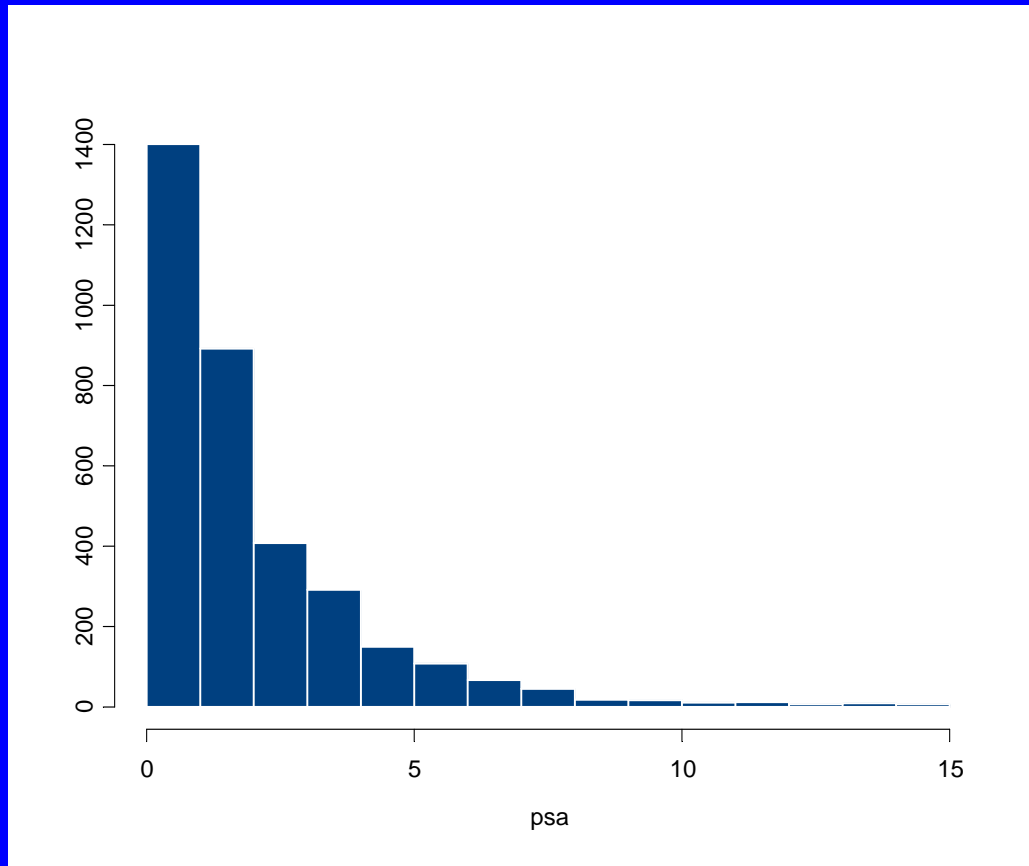
# Example: Morgan's PSA Data

---

- ◆ NEJM 1996;335:304
- ◆ 3475 without prostate cancer
- ◆ PSA grouped into units of ng/ml

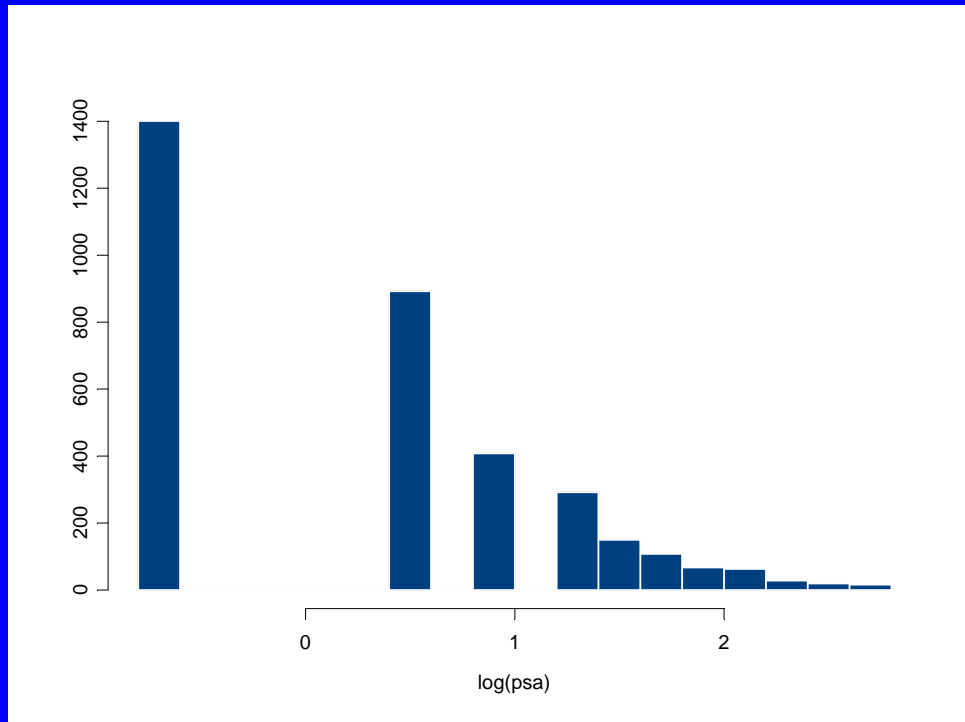
# Morgan's PSA Data

---

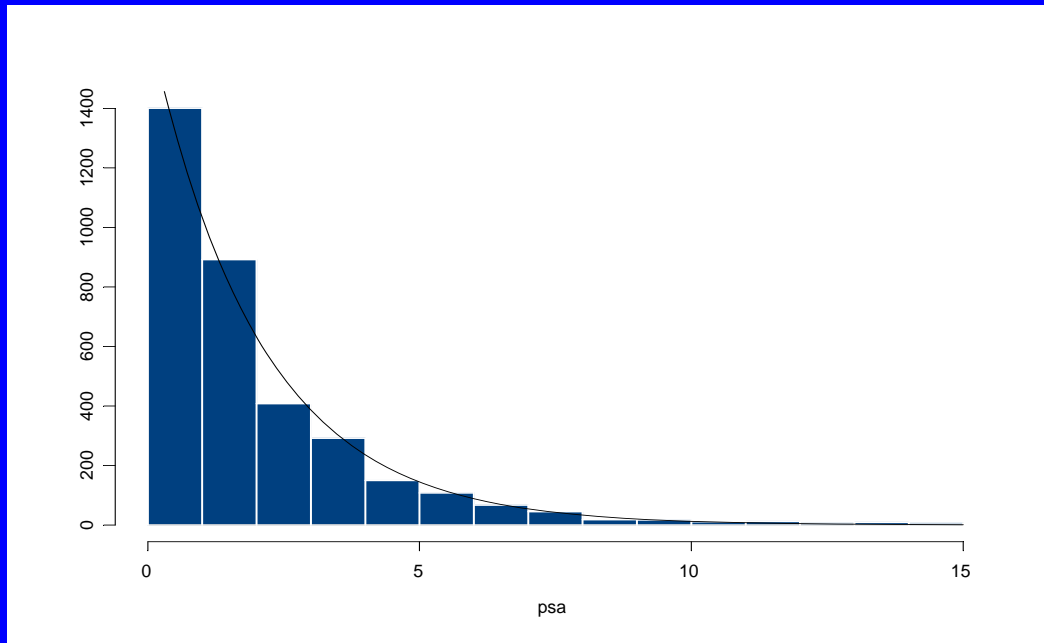


# Morgan's PSA Data

---



# Morgan's PSA Data-Exponential Distribution



## Morgan's PSA Data-Exponential Distribution

---

- ◆ Thus, PSA follows an exponential F, and this implies that the upper limit is obtained as 3 times the mean:
- ◆ 95% upper reference limit  $\sim 6.1$  ng/ml
- ◆ Lower limit is 0.

# Exponential Distribution Function

---

- ◆ Many clinical variables begin at 0.
- ◆ For many the most likely values are also at or just above 0.
- ◆ For such variables the exponential distribution function may be appropriate.

# The Exponential Distribution Function

---

- ◆  $f(x) = a * \exp(-a * x)$
- ◆  $F(x) = 1 - \exp(-a * x)$
- ◆ mean =  $1/a$
- ◆ 95% limits: 0 to  $3 * \text{mean}$

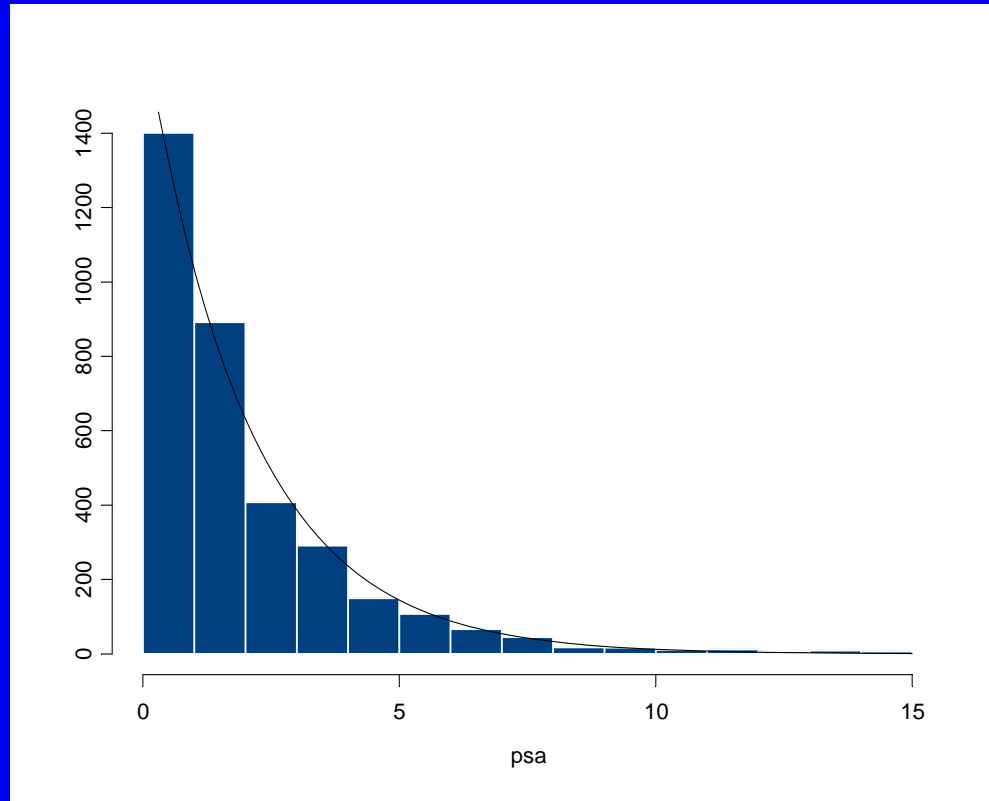
# The Exponential Distribution Function

---

- ◆ For any clinical lab variable,  $x$
- ◆ Plot the frequency distribution of  $x$
- ◆ Estimate  $a$  as  $1/\text{mean of } x$
- ◆ Overlay plot of  $n * f(x) = a * \exp(-a * x)$  and see how the two plots compare

# The Exponential Distribution Function

---



# Reference Limits-Non-Parametric

---

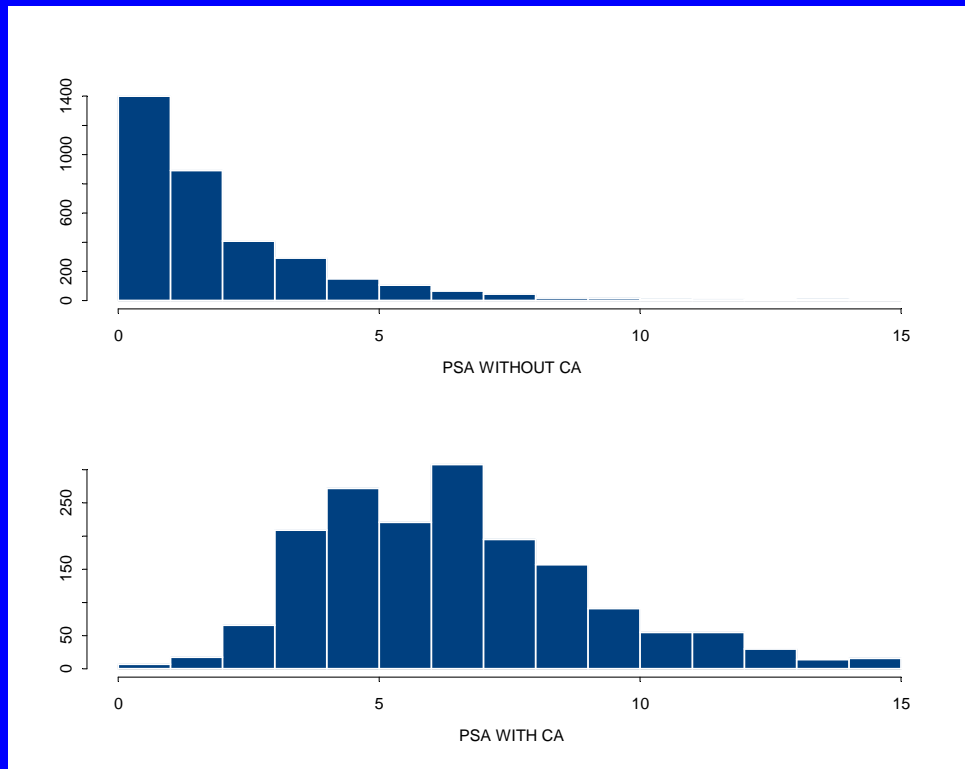
- ◆ Order the data by increasing  $x$ .
- ◆ Count off 2.5%; the  $x$  value at this point is the lower reference limit,  $x_l$ .
- ◆ Count off 97.5%; the  $x$  value at this point is the upper reference limit,  $x_u$ .
- ◆ For Morgan data 95% upper reference limit is approximately 5.9 ng/ml.

# The problems with reference limits

---

- ◆ The approach is univariate.
- ◆ i.e. it considers just the reference population, not the diseased populations.
- ◆ It deals with  $P(x \text{ outside range} | \text{Ref. Pt.})$ , but it ignores  $P(x | \text{Diseased Pt.})$ .
- ◆ It ignores the influence of multiple variables.

# Distributions of PSA in Benign and Ca Populations (Morgan Data)



# Dealing with the Diseased Group: Use of 2 x 2 Table

---

- ◆ Key to many laboratory concepts such as positive predictive value, sensitivity, specificity, true positive, false negative, etc.
- ◆ Starting point is two event sets:
- ◆ Disease events:  $\{D+, D0\}$
- ◆ Test events:  $\{T+, T0\}$
- ◆ Presume that T is a potential test for D.

# 2 X 2 Table

---

◆	T+	T0
◆ D+	True positives	False negatives
◆ D0	False positives	True negatives
◆ and Total	= n	

# 2 X 2 Table For D and T

---

- ◆            T+    T0
- ◆ D+        a     b
- ◆ D0        c     d            and  $a+b+c+d = n$
- ◆ a = True positives
- ◆ b = False negatives
- ◆ c = False positives
- ◆ d = True negatives

# Conditional Probabilities

---

- ◆ Sensitivity =  $P(T+|D+) \sim a/(a+b)$
- ◆ Specificity =  $P(T0|D0) \sim d/(c+d)$
- ◆ Pos Predictive Value =  $P(D+|T+) \sim a/(a+c)$
- ◆ Neg Predictive Value =  $P(D0|T0) \sim d/(b+d)$

# Conditional Probabilities for Disease and Test. Depends on What's On Top

---

- ◆ Test Result on Top:
  - ◆  $P(T+|D+)$  &  $P(T0|D0)$  i.e. sensitivity & specificity
- ◆ Disease on Top:
  - ◆  $P(D+|T+)$  &  $P(D0|T0)$  i.e. PPV & NPV

# Conditional Probabilities for Disease and Test

---

- ◆ For Sensitivity & Specificity, one collects pts with either D+ or D0 and records the frequency of T+ and T0 and :
- ◆  $P(T+|D+) \sim \text{no. } T+ \text{ \& } D+ / \text{no. } D+$
- ◆  $P(T0|D0) \sim \text{no. } T0 \text{ \& } D0 / \text{no. } D0$

# Conditional Probabilities for Disease and Test

---

- ◆ For PPV & NPV, one collects pts with either T+ or T0 and records the frequency of D+ and D0 and :
- ◆  $P(D+|T+) \sim \text{no. } D+ \text{ \& } T+ / \text{no. } T+$
- ◆  $P(D0|T0) \sim \text{no. } D0 \text{ \& } T0 / \text{no. } T0$

# Example: Babaian PSA Data

---

	PSA > 4	<4
◆ CA	48	21
◆ B9	58	277
◆ PPV	$= 48/106 = 0.45$	
◆ NPV	$= 277/298 = 0.93$	
◆ Sens	$= 48/69 = 0.70$	
◆ Spec	$= 277/335 = 0.83$	

# Conditional Probabilities

---

- ◆ In general positive predictive value, PPV or  $P(D+|T+)$ , is a useful result for pt and clinician, because it answers the pt's question what is the chance of my having disease, given this lab test result.

# Conditional Probabilities

---

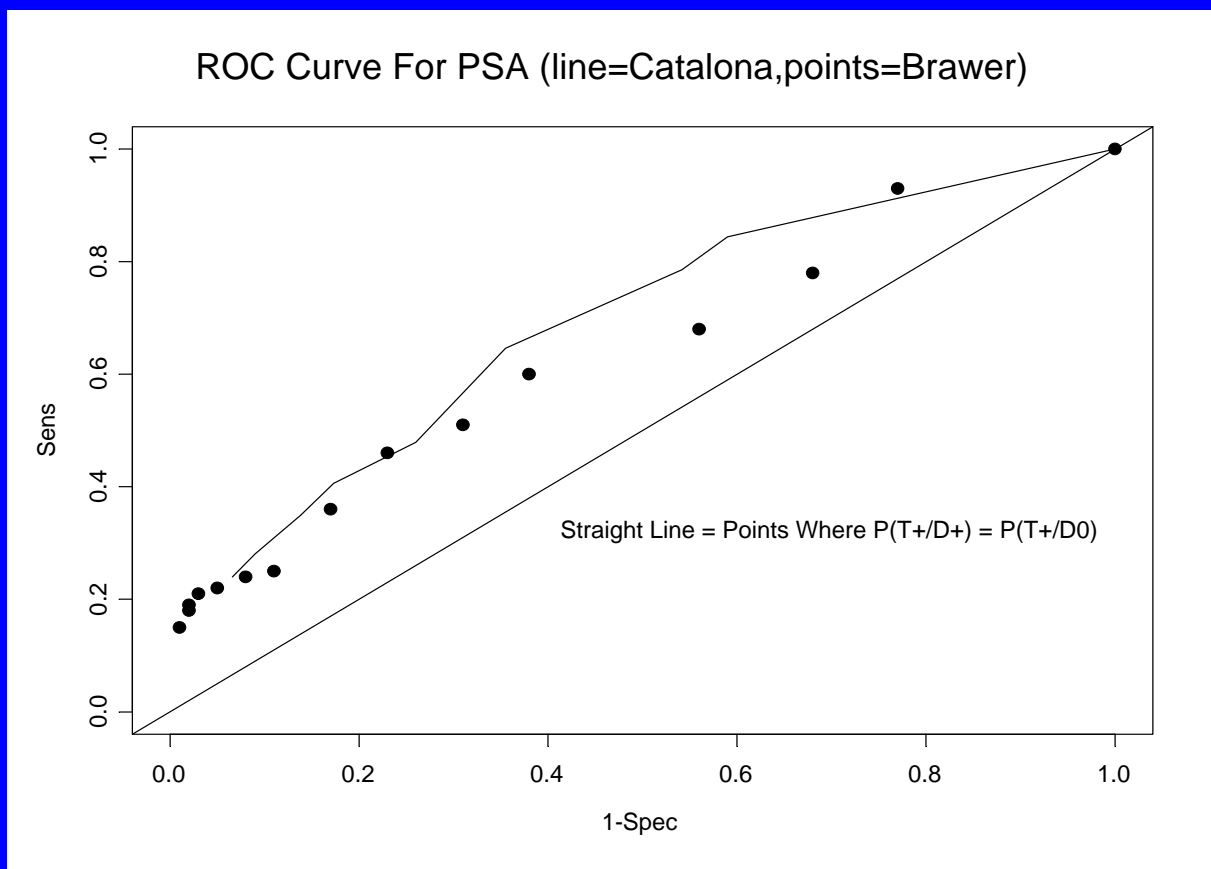
- ◆ Predictive value of negative test also important, i.e.  $P(D_0|T_0)$ , because it answers the pt's question what is the chance of my not having disease given a negative test result.

# Receiver Operating Characteristic (ROC) Curves

---

- ◆ Some more recently have termed it “relative operating characteristic” curve
- ◆ Plot of sensitivity (  $P(T+|D+)$  ) on vertical
- ◆ vs 1-specificity (i.e.  $P(T+|D0)$  ) on horizontal
- ◆ Here, as before, T and D are binary.
- ◆ Thus, the ROC is a plot of P(True Pos) vs P(False Pos)

# ROC for PSA

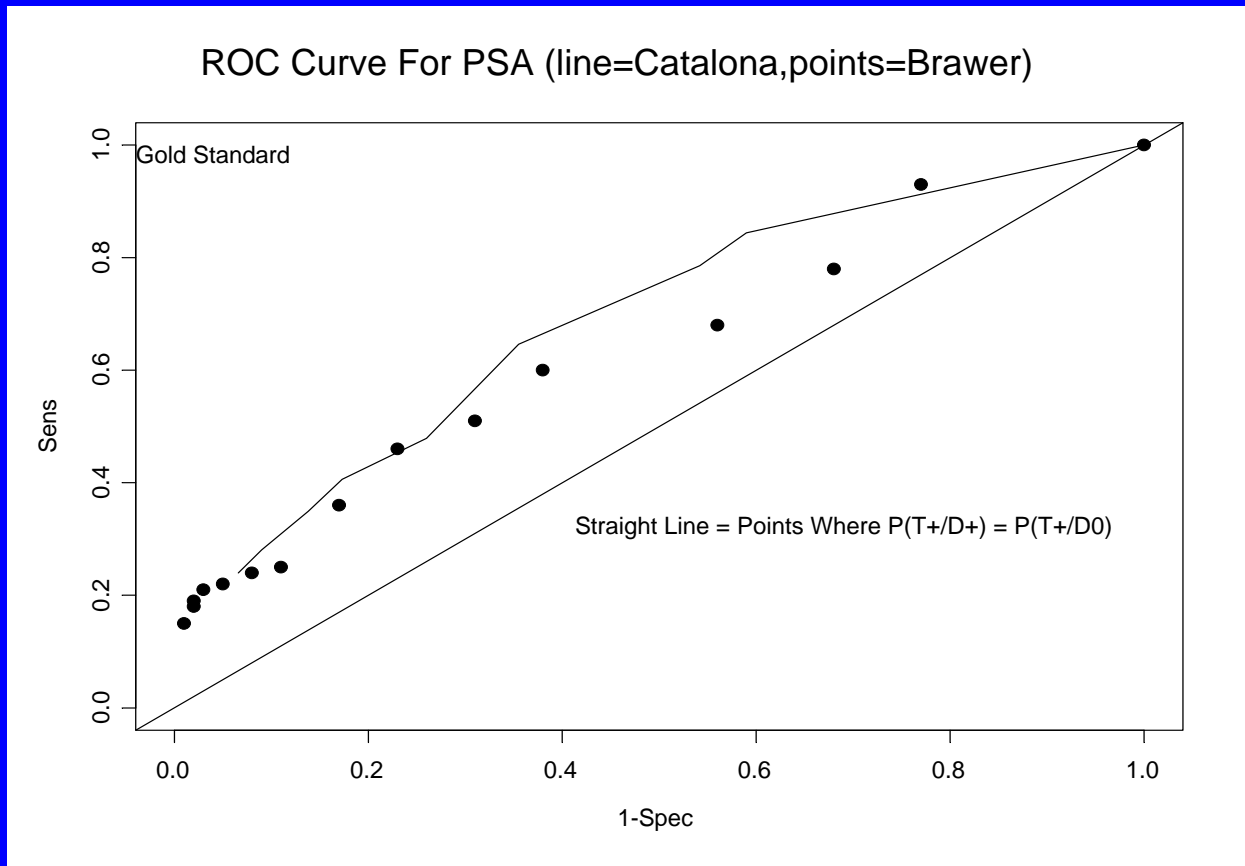


# ROC Curve

---

- ◆ Gold standard test result:  $P(T+|D+)=1$ ;  $P(T+|D0)=0$ , i.e. a single point at  $(x=0, y=1)$ .
- ◆ The straight line shows where  $P(T+|D+) = P(T+|D0)$ , i.e. where  $P(\text{True } +) = P(\text{False } +)$ .
- ◆ At this line the  $\text{Odds}(D+|T+) = \text{Odds}(D+)$ , i.e. the post test odds are the same as pretest odds (test is worthless).

# Gold Standard: Sens=1; 1-Spec=0

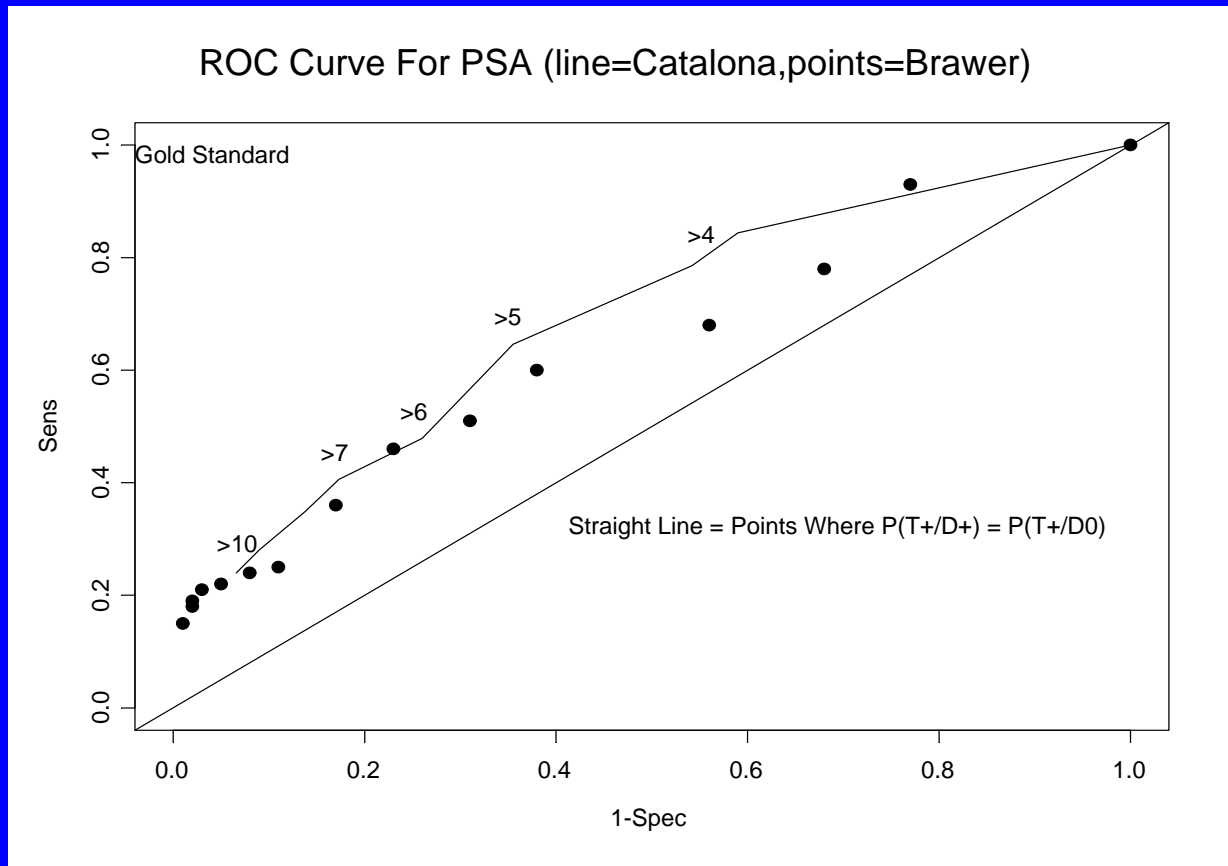


# ROC: What makes the curve?

---

- ◆ If a test gives rise to a single sensitivity and a single specificity, then it should be represented by a single point on the ROC.
- ◆ Multiple sensitivity-specificity pairs or points on the ROC are often due to different cutpoints for a continuous  $x$ .
- ◆ Example: different cutpoints for PSA.

# ROC for Cutpoints of PSA



# ROC-What makes the curve?

---

- ◆ Using the ROC implies one has to transform a continuous test variable into a binary one, i.e. one that is + or 0.
- ◆ E.g.  $>4$ ,  $>5$ , ...
- ◆ The best cutpoint is the one closest to the Gold Standard.

# ROC--Good Curves=Good Tests

---

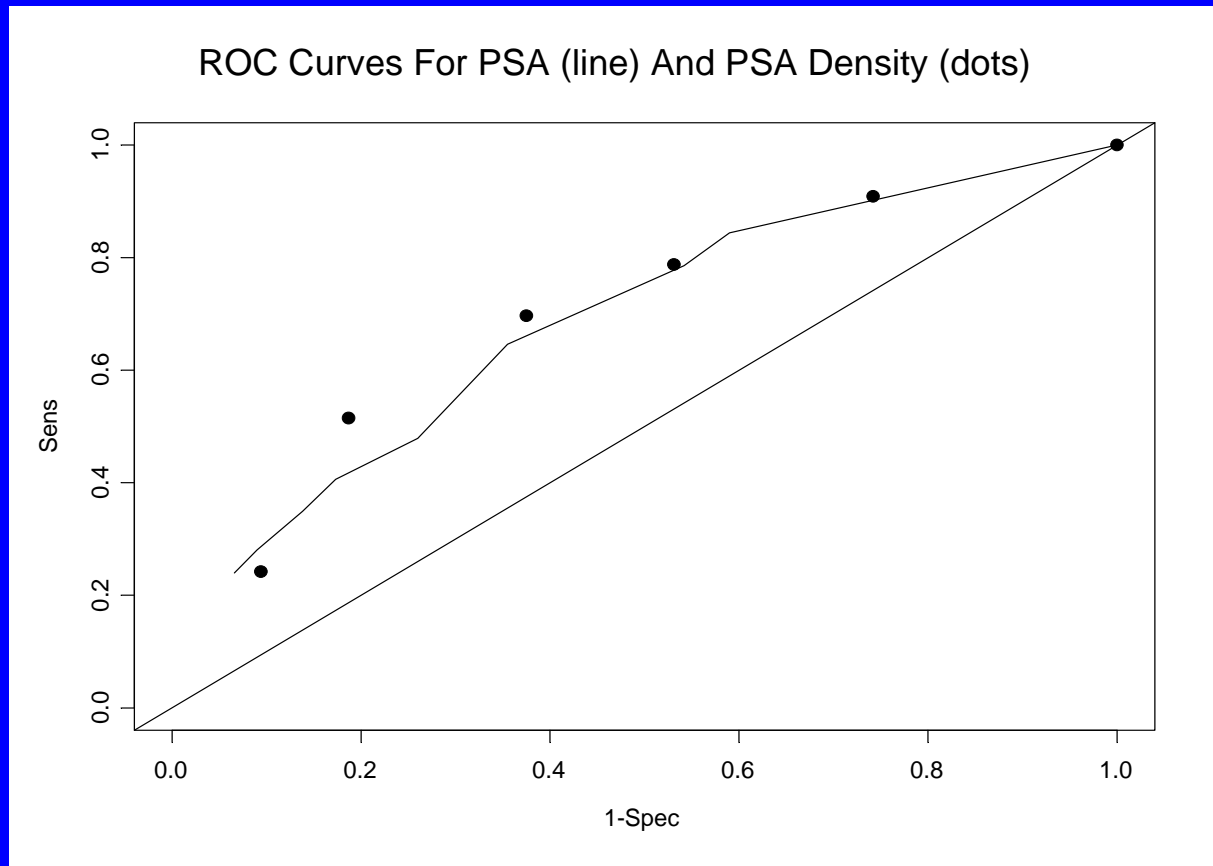
- ◆ The closer the curve is to the gold standard ( $x=0, y=1$ ) the better the diagnostic test.
- ◆ The further the curve is from the equal probability line, the better the test.
- ◆ The larger the area under the ROC curve, the better the test.
- ◆ Thus, one use of ROC is to compare tests.

# Using ROC to Compare 2 Tests

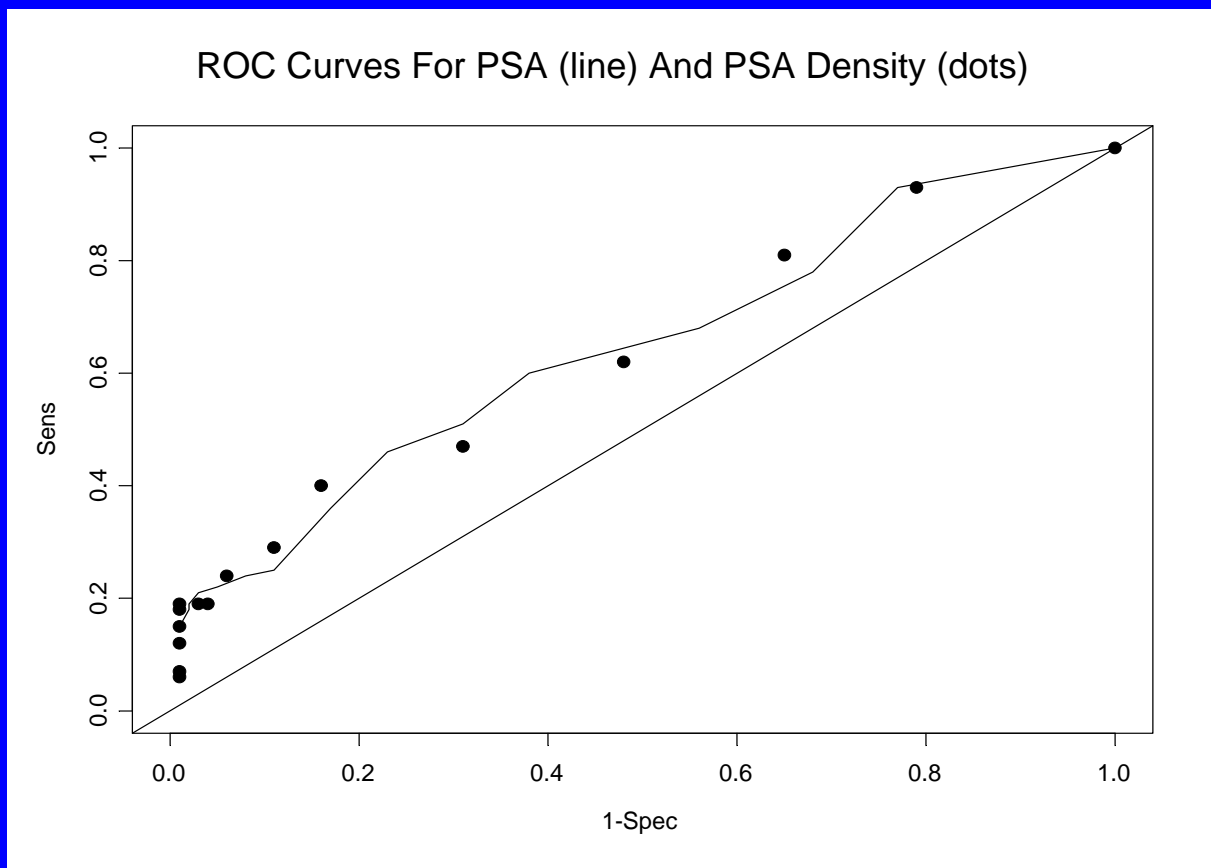
---

- ◆ PSA is known to increase with the size of the prostate.
- ◆ PSA density is a correction of PSA for gland volume as measured by US.
- ◆  $\text{PSA density} = \text{PSA} / \text{Gland Vol}$
- ◆ How do their ROC compare?

# Comparing PSA vs. PSA Density: Catalona Data



# Comparing PSA vs. PSA Density: Brawer Data



# ROC--Comparing 2 Tests

---

- ◆ By this simple graphical technique, PSA density seems no better than PSA alone.
- ◆ Critique: In order to use ROC one must have a single continuous variable that is broken into cutpoints. PSA density achieved this by making a ratio, and this forces PSA and US volume to have the same weight.

# Importance of D0 Population

---

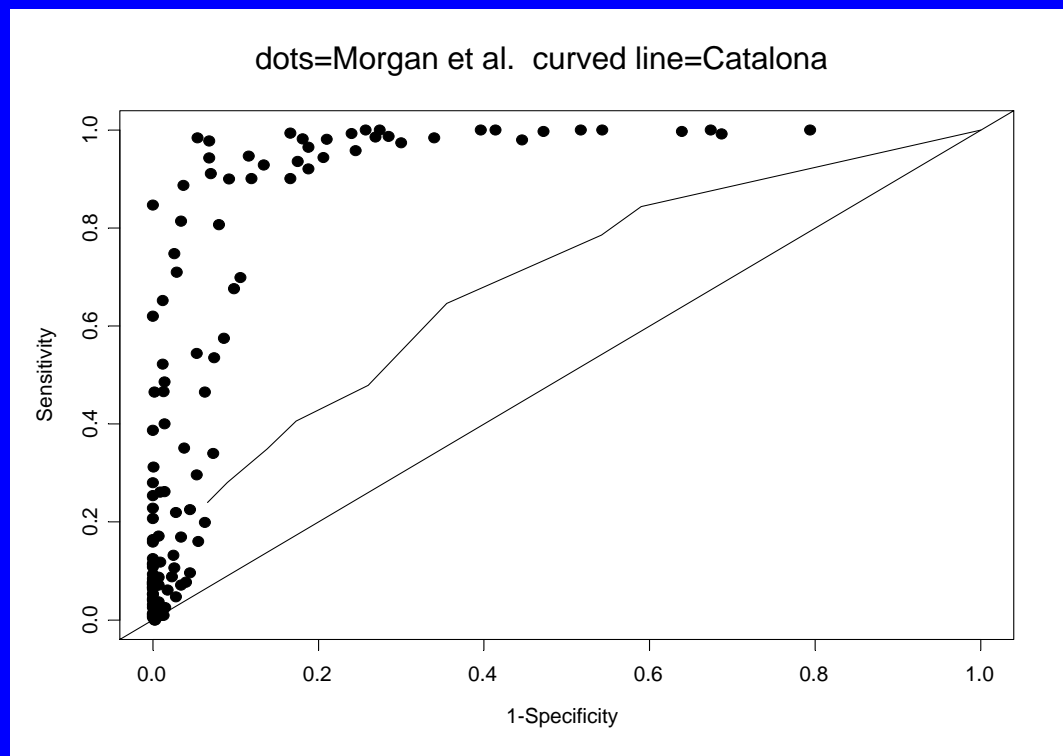
- ◆ When estimating the sensitivity and specificity, the D0 population used is critical. Make sure it is the reference you want to use.
- ◆ Catalonia: D0 was all with a negative bx, DRE-, or low PSA.
- ◆ Morgan: D0 was all with negative bx + all who had no CA during follow-up.

# ROC Curve

---

- ◆ Remember the ROC curve is the plot of sensitivity or  $P(T+|D+)$  against
- ◆ 1-specificity or  $P(T+|D0)$

# Morgan et al. PSA Data



# Issue of Cutpoints in a Continuous Test Result, $x$

---

- ◆ When all  $x$  results are transformed into negative ( $x < a$ ) or positive ( $x \geq a$ ), then information is lost.
- ◆ Alternative is to calculate the PPV for a specific value of  $x$ , i.e.  $P(D|x \geq a)$ .
- ◆ Logistic regression can do this and also incorporate other key variables.

# ROC Curves

---

- ◆ The original use of the ROC was to evaluate observers' recognition of military aircraft during WWII.
- ◆ Thus, points on the ROC were different observers, each with a unique pair of sensitivities and specificities.
- ◆ Obvious application to AP:

# ROC for Anatomic Pathology

---

- ◆ Whenever one can record sensitivity and specificity for a binary test,  $T$ , and a binary outcome  $D$  (need not be a disease), then one could apply the ROC.
- ◆ e.g. Let  $D$  be determined by biopsy and  $T$  be a positive or negative cytology,
- ◆ then points on the ROC could be different cytopathologists or cytotechnicians.

# ROC for Urine Cytopathology

---

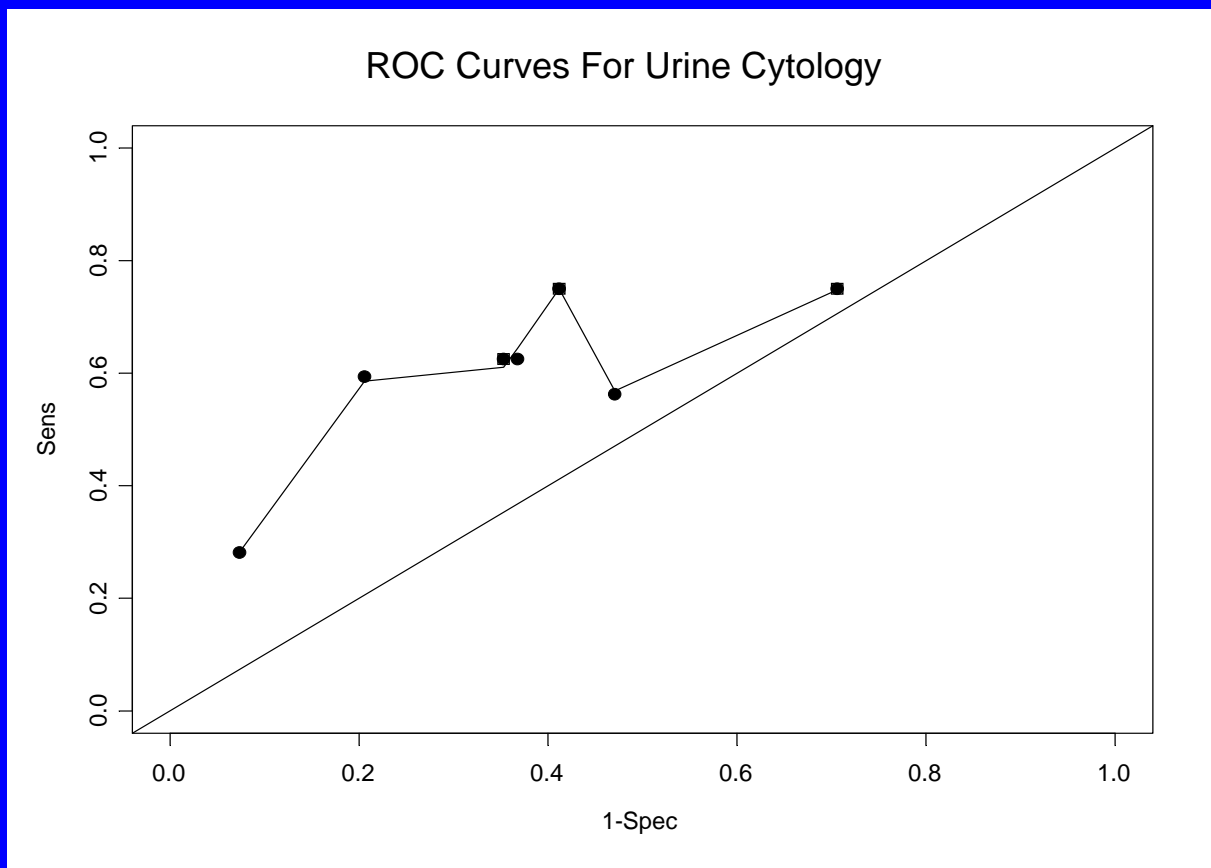
- ◆ van der Poel et al data (Mod Pathol 1997;10:976-982).
- ◆ 4 Experts, 100 Bladder Washings, All cystoscoped but not all biopsied
- ◆ Gold Standard: Cystoscopy

# ROC for Urine Cytopathology

---

- ◆ Three categories of cytological diagnosis:
- ◆ Negative, atypical, tumor.
- ◆ For each observer and each threshold of a positive cytology the no. of cases that were D+T+, D+T0, D0T+, & D0T0 recorded.
- ◆ Sensitivities and specificities were calculated as in the 2 x 2 table.

# ROC for Urine Cytology



# ROC for Urine Cytopathology

---

- ◆ 3 Points were close to non-ideal line:
- ◆ pathologist 4 for any threshold, pathologist 3 with positive for tumor threshold, & pathologist 1 with atypia as threshold.

# Bayes Theorem

---

- ◆  $P(D+ \& T+)$
- ◆  $P(D+|T+) = \frac{\text{-----}}{P(T+)}$
- ◆  $P(T+|D+) * P(D+)$
- ◆  $PPV = \frac{\text{-----}}{P(T+)}$
- ◆  $P(T+)$

# Bayes' Theorem

---

$$\begin{aligned} & \text{Sens} * P(D+) \\ \text{PPV} &= \frac{\text{Sens} * P(D+)}{P(T+|D+) * P(D+) +} \\ & P(T+|D0) * (1-P(D+)) \end{aligned}$$

◆  $P(T+ | D0)$  is the false pos. prob. (FP)

# Bayes' Theorem

---

◆ Thus,

◆  $\text{Sens} * P(D+)$

◆  $PPV = \frac{\text{Sens} * P(D+)}{\text{Sens} * P(D+) + FP * P(D0)}$

◆  $\text{Sens} * P(D+) + FP * P(D0)$

◆ Dividing by  $\text{Sens} * P(D+)$  leads to :

# Bayes' Theorem

---

◆ Finally,

$$\begin{aligned} & \text{◆} \quad \frac{1}{\text{◆}} \\ & \text{◆} \quad \text{PPV} = \frac{1}{1 + \frac{\text{FP} * \text{P(D0)}}{\text{Sens} * \text{P(D+)}}} \end{aligned}$$

# Bayes' Theorem for Prostate Ca

---

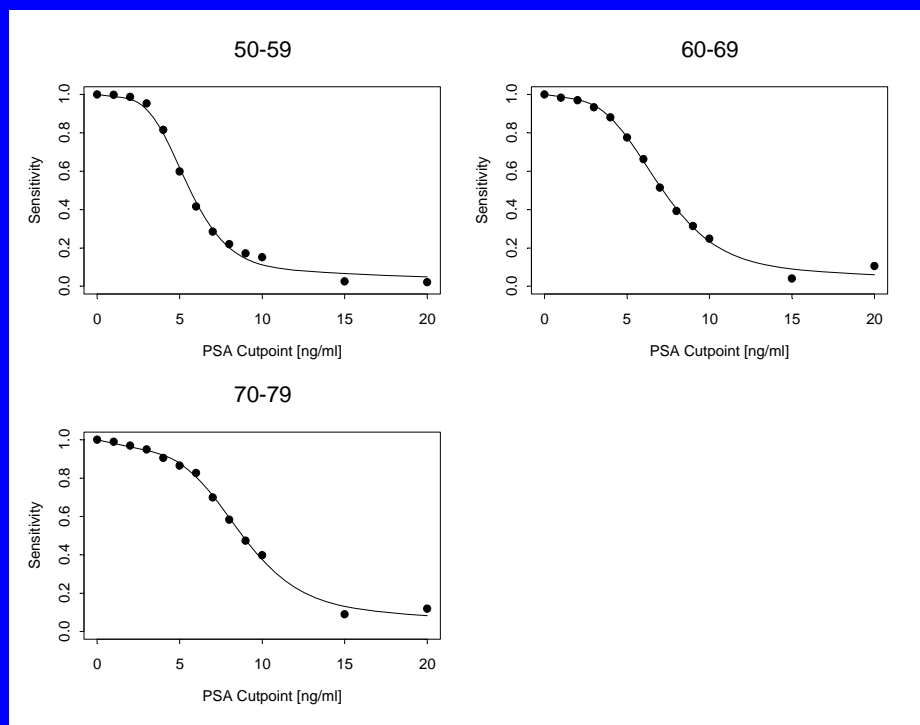
- ◆ PPV for prostate cancer:  $P(\text{Ca} \mid \text{PSA} > x)$
- ◆ Sensitivity for prostate cancer:
  - ◆  $P(\text{PSA} > x \mid \text{Ca})$
- ◆  $P(\text{Ca})$  for prostate cancer is the cumulative incidence
- ◆ FP for prostate cancer:  $P(\text{PSA} > x \mid \text{B9})$

# Bayes' Theorem for Prostate Ca

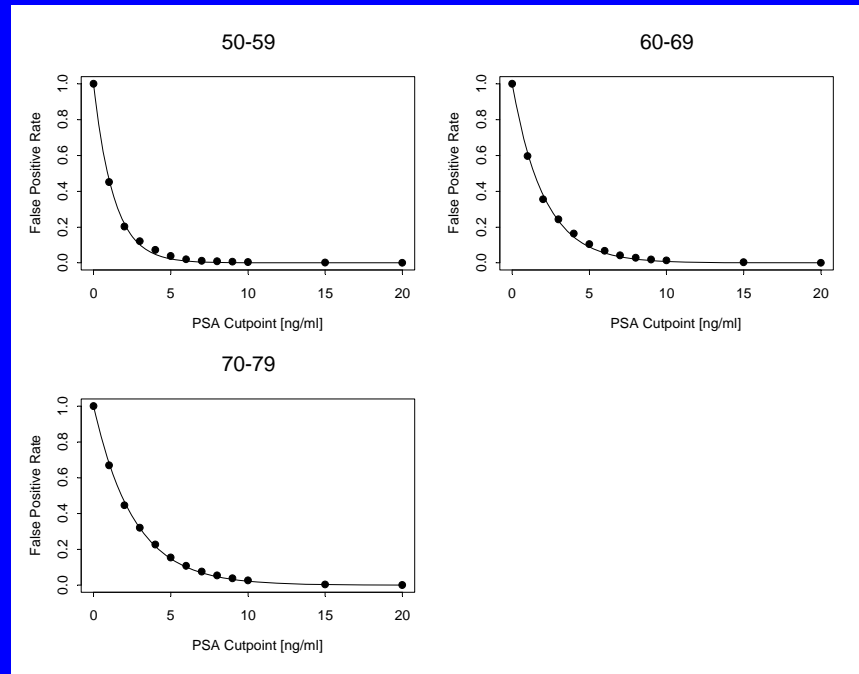
---

- ◆ Sensitivity:  $>2700$  men with Ca,  $F =$  sum of gamma and exponential probability functions
- ◆ FP:  $> 99,000$  men without Ca,  $F =$  exponential probability function
- ◆ 3 Age groups: 50-59, 60-69, 70-79
- ◆  $P(\text{Ca}) =$  cumulative incidence from SEER

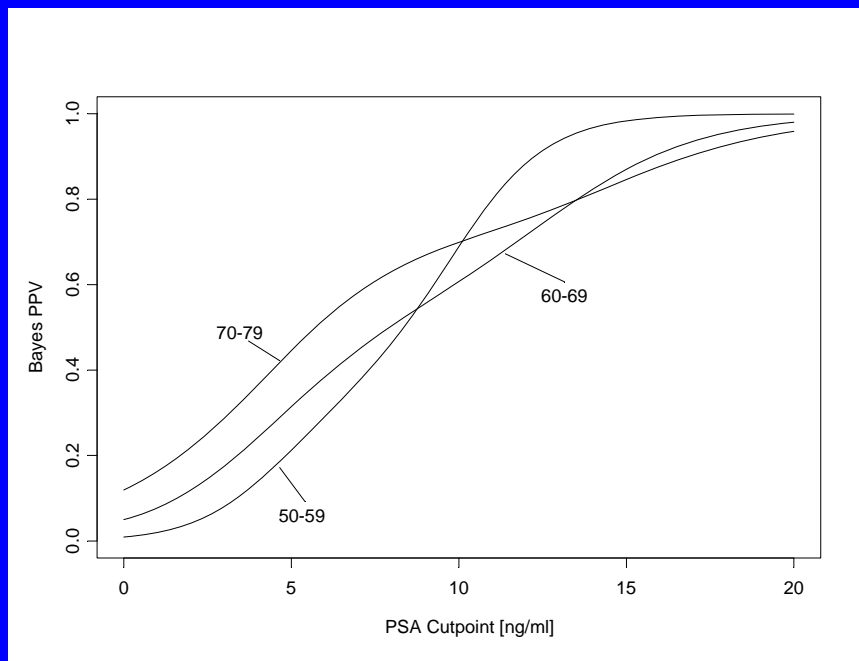
# Sensitivities for Prostate Ca



# FP for Prostate Ca



# Bayes' PPV for Prostate Ca



# Bayes' PPV for Prostate Ca

---

- ◆ PPV for prostate Ca reflects: PSA, age of pt., and overall age-specific incidences of Ca in complex and overlapping fashion.
- ◆ Lowering the cutpoint for PSA in younger men should result in many negative biopsies.

# Introduction to Statistical Testing

---

- ◆ Is the value of serum PSA statistically independent from the presence of prostate cancer?
- ◆ i.e. does the test, serum PSA, relate to the presence of the disease, prostate cancer?
- ◆ i.e. are T and D associated?

# Are D and T Statistically Independent?

---

- ◆ If so, then T is not useful for the diagnosis of D.
- ◆ From the rules for probability we know that two events like D+ and T+ are independent iff  $P(D+\&T+) = P(D+) * P(T+)$

# Are D and T Statistically Independent?

---

- ◆ We begin by assuming that D and T are independent (the Null hypothesis).
- ◆ The observed no. of pts with D+ & T+ is “a” in the 2 x 2 table.
- ◆ The expected no. by the Null is
- ◆  $n * P(D+) * P(T+)$

# The Chi-square Test for Statistical Independence

---

- ◆ The chi-square statistic is the sum of
- ◆  $[(\text{observed}-\text{expected})/\text{expected}]^2$
- ◆ It has 1 degree of freedom.

# Degrees of Freedom

---

- ◆ In general degrees of freedom relate to the amount of data one has, decreased by the number of things estimated from the data.
- ◆ Here we had 4 entries in the 2x2 table; subtracting a known total of  $n$ , and 2 estimates of  $P(D+)$  and  $P(T+)$ , we are left with 1 d.f.

# Are D and T Statistically Independent?

---

- ◆ Then if the obs and exp are very different, the chi-square statistic is large, the probability for this result in the chi-square table is small (i.e. the p value), so we reject the Null hypothesis that D and T are independent.
- ◆ Alternatively, if obs and exp are close, then the chi-sq is small and the p large.

# Example: Babaian '92 PSA Data

---

- ◆ PSA > 4 < 4 ng/ml
- ◆ CA 48 21
- ◆ B9 58 277 n = 404
- ◆ Compare:
  - ◆ 48 vs. 404 (69/404) (106/404)
  - ◆ 21 vs. 404 (69/404) (298/404)
  - ◆ 58 vs. 404 (335/404) (106/404), etc.

# Example: Babaian '92 PSA Data

---

- ◆ Subtract, square, and sum these differences.  
The result is a Chisquare value of 78
- ◆ The p value is approximately 0.
- ◆ Thus, the test,  $PSA > 4$ , and presence of prostate CA are unlikely to be independent from one another.

# Statistical Tests of Null Hypothesis

---

- ◆ Assume the Null, i.e. independence of D & T.
- ◆ Obtain value for statistic  $s$ .
- ◆ Plug  $s$  into the distribution function  $F(s)$  to obtain a  $p$  value.
- ◆  $P$  value = the probability that the observed value is  $\geq s$ , given that the null is true.

# p Value

---

- ◆ P value =  $P(s \mid \text{Null})$
- ◆ When the p value is low, then the Null hypothesis is likely to be false.

# Examples of Null

---

- ◆ 2 Events are Independent
- ◆ Proportions of 2 populations are equal.
- ◆ Means of 2 populations are equal.
- ◆ Coefficient in regression =0.
- ◆ Survival of 2 groups are equal.

# Type I Error

---

- ◆ Type I error occurs when we reject the Null because of low p value, but the Null is actually true.
- ◆ The p value is also the probability of making a Type I error.

# Type II Error--The Null is False

---

- ◆ Error is that we accept the Null, because the p value is too high.
- ◆ Type II errors are due to small data sets or uncontrolled variance in the data.
- ◆ Warning: with small numbers the statistics,  $s$ , may not follow  $F(s)$ , i.e. you can't have that p value!

# Type II Error--The Null is False

---

- ◆ Probability of Type II error is often denoted:  $\beta$
- ◆ Power is  $1 - \beta$ , i.e. the probability of not making a Type II error.
- ◆ The power relates directly to the size of data set as well as to the type of statistical test used for the Null.

# Sample Sizes

---

- ◆ In general 100's are good (NCCLS: 120 per group).
- ◆ <100 tricky but 50-100 doable;  $\leq 20$  weak
- ◆ For multivariable analyses, one needs approximately 15 cases for each covariate.
- ◆ Small effects on uncommon events require huge samples and long follow-up.

# In Summary

---

	Truth About the Null	
◆ Action	Null True	Null False
◆ Accept Null	OK	Type II
◆ Reject Null	Type I	OK

- ◆  $P(\text{Type I}) = p \text{ value}$
- ◆  $1 - P(\text{Type II}) = \text{power}$

# Type III Errors (informal)

---

- ◆ p value  $< 0.05$ , Null rejected, but results are not important, because:
- ◆ The variables explain too little of the variance in the data to be useful.
- ◆ Other variables are more important, have bigger effect and are neglected in this study.
- ◆ Model won't validate with new data.
- ◆ New technology too expensive, too difficult.

# Cookbook Approach to Statistics

---

- ◆ Examine the random variables.
- ◆ Dependent and explanatory.
- ◆ Discrete? Continuous?
- ◆ Paired observations?
- ◆ Univariate vs. multivariate approach?
- ◆ See table and text in hand-out for further information.

# Chi-Square Test of Statistical Independence

---

- ◆ Two discrete random variables
- ◆ e.g. D and T
- ◆ e.g. Prostate Ca. and PSA with a cutpoint
- ◆ Non-parametric in that the two variables need not follow a particular distribution function.

# FNA of Thyroid Lesions

---

- ◆ Ersoz et al. paper: Cancer Cytopathol. 2004;102:302-307.
- ◆ Cellularity Scored: 1 - 4
- ◆ Microfollicles: 0, 1-25, 50-75, >75%
- ◆ Nuclear Diameter Scored: 1 - 3
- ◆ Nuclear Pleomorphism Scored: 0 - 2

# FNA of Thyroid Lesions

---

◆ Neg.	Tumor-B9	Ca
◆ 63	99	62

- ◆ Authors provided sensitivities, specificities, PPV, and NPV for a malignant dx. each of the four variables.

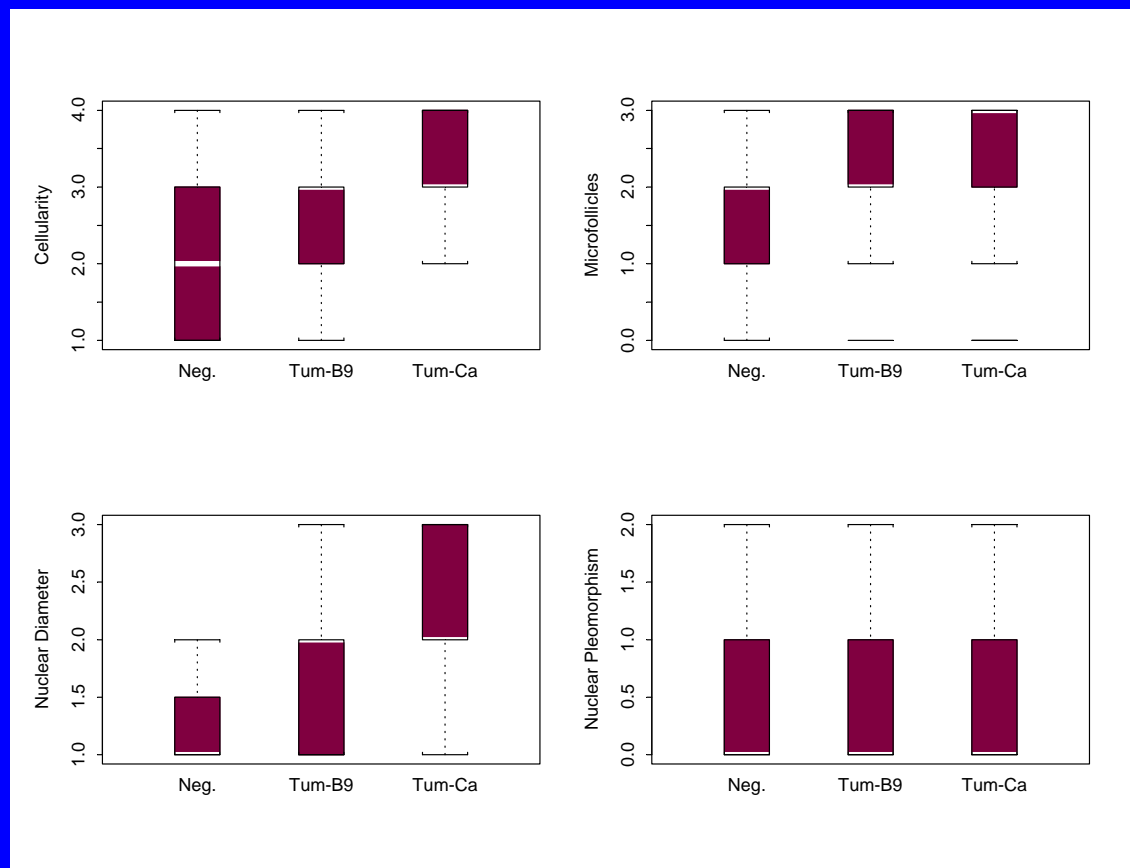
# FNA of Thyroid Lesions

---

◆ Var	Sens	Spec	PPV
◆ Cell.	78.5	50	34.3
◆ Mfol.	80	29.2	29.2
◆ Ndiam.	76.9	55.8	34.4
◆ Npleo.	45	61.1	39.1

◆ No statistical tests were done.

# FNA of Thyroid Lesions



# Kruskal-Wallis Non-Parametric Test

---

- ◆ Based on ranks of the dependent variable,  $y$ .
- ◆ Does not assume any distribution function for  $y$ .
- ◆ Tests whether  $y$  differs among categories of a second variable,  $x$ .
- ◆ Here, e.g.  $y$  is cellularity, etc. and  $x$  is  $dx$ .

# FNA of Thyroid Lesions

---

◆ Results from Kruskal-Wallis test

◆ Variable	p value
------------	---------

◆ Cellularity	0.04
---------------	------

◆ Microfollicles	0.07*
------------------	-------

◆ Nuc. Diameter	0.003
-----------------	-------

◆ Nuc. Pleomorphism	0.6
---------------------	-----

◆ \*  $p=0.02$  for Neg. vs. any tumor

# FNA of Thyroid Lesions

---

- ◆ Conclusions:
- ◆ Biggest differences were in nuclear diameter (i.e.  $< 1$  rbc,  $= 2$  rbc,  $> 2$  rbc)
- ◆ Cellularity and number of microfollicles also important.
- ◆ Nuclear pleomorphism does not seem important.

# Multivariate Statistical Tests

---

- ◆ Chi-square test and Kruskal-Wallis test are examples of tests involving one dependent  $y$  variable and one explanatory  $x$  variable.
- ◆ What is more often needed are tests that allow multiple explanatory variables, i.e.  $x_1, x_2, x_3, \dots$

# Linear Regression Models

---

- ◆ Continuous dependent variable:  $y$
- ◆ Continuous or discrete explanatory variables:  $x_1, x_2, x_3, \dots$
- ◆ Model:
- ◆  $y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + \dots +$   
error

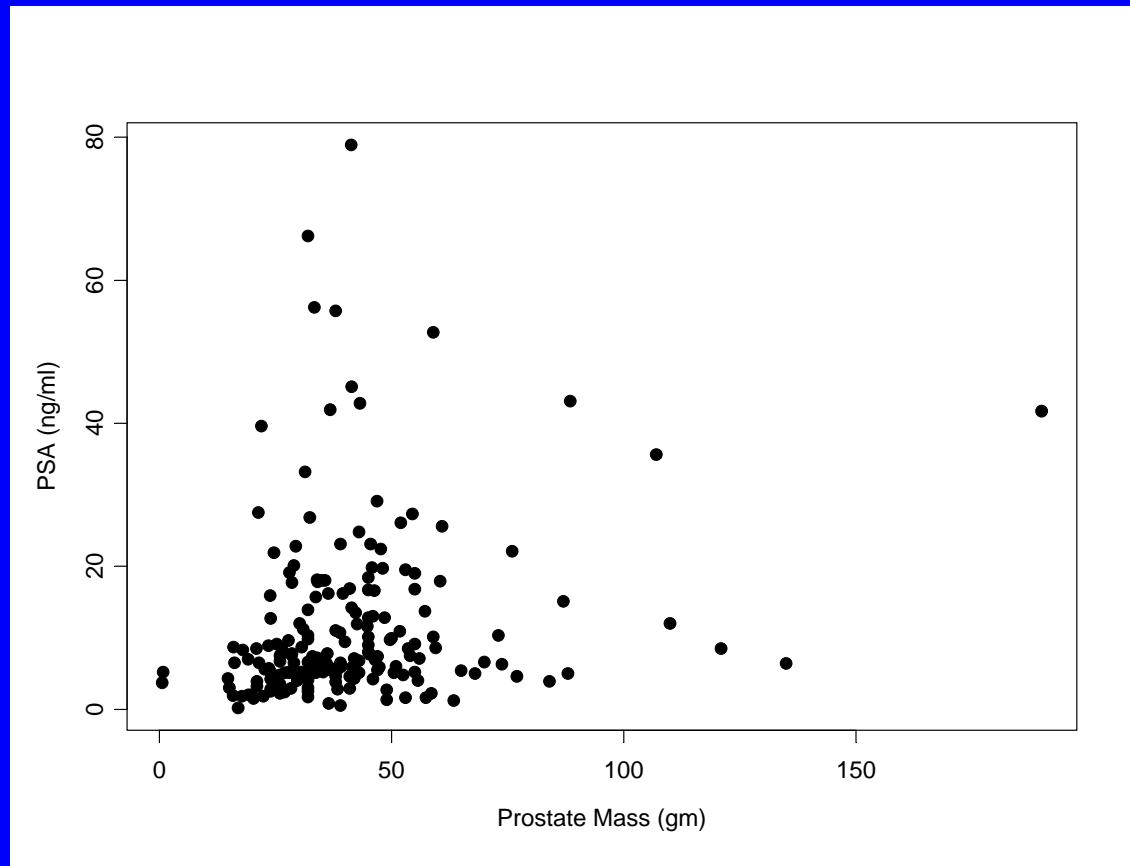
# Linear Regression Models

---

- ◆ Linear regression model estimates values for  $b_0$ ,  $b_1$ ,  $b_2$ , ...
- ◆ error term is assumed to follow normal distribution function
- ◆ Null hypothesis is that particular  $b$  coefficients equal 0 (e.g. if  $b_1=0$ , then  $x_1$  makes no contribution)

# Example Serum PSA vs. Mass of Prostate Gland

---



# Serum PSA and Prostate Gland

---

- ◆ Theoretical analyses suggest that PSA should depend directly on the mass of the prostate.
- ◆ Linear regression with one variable:
- ◆  $PSA = 0.246 * \text{mass}$  (t~12, p~0)
- ◆ Every gm of prostate leads to 0.246 ng/ml of serum PSA

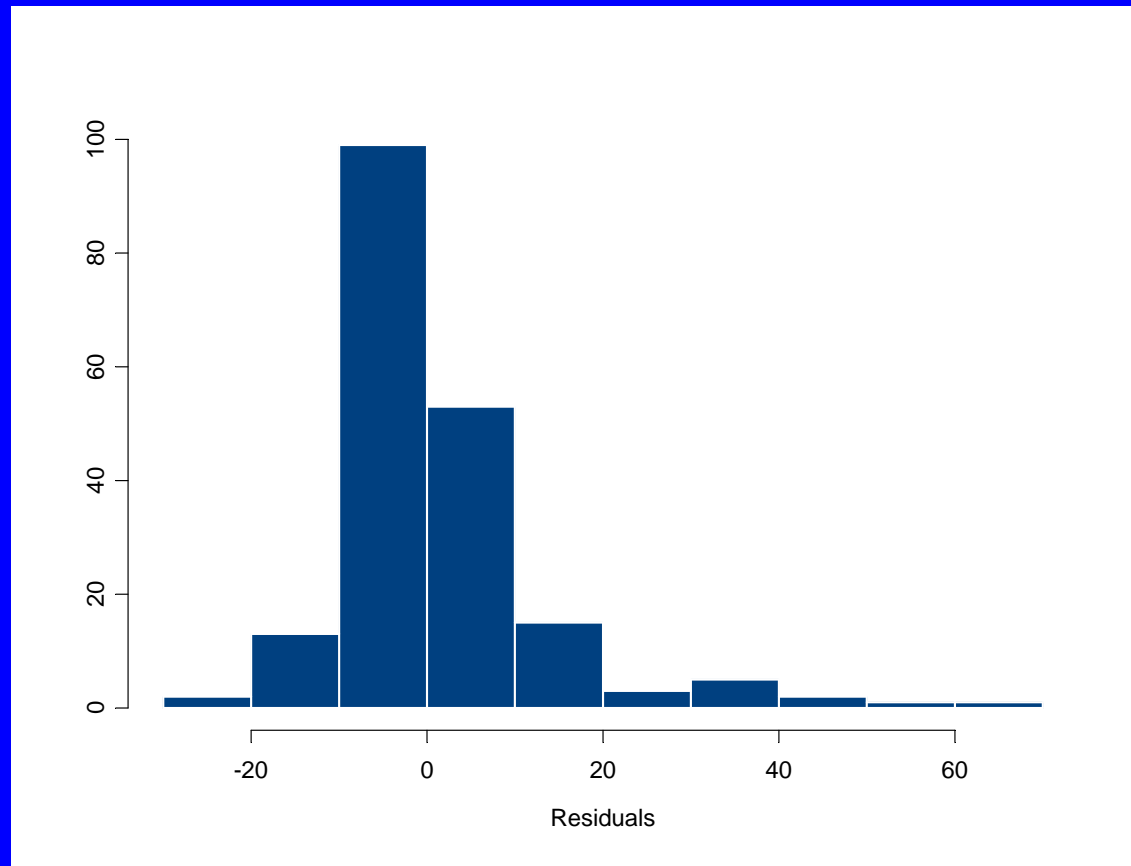
# Serum PSA and Prostate Gland

---

- ◆ To use linear regression:
- ◆  $y$  must be continuous
- ◆  $y$  need not be normally distributed
- ◆ BUT the residuals must be approximately normally distributed
- ◆ e.g. residual = observed PSA - PSA predicted by linear model

# Residuals of Linear Model for PSA

---



# Serum PSA and Prostate Gland

---

- ◆ Residuals ranged from -26 to 68, and regression R-square was 0.46, i.e. just 46% of the noise was explained by the linear model.
- ◆ Other variables important?

# Serum PSA and Prostate Ca.

---

◆ Linear Model with 2 Explanatory Variables:

◆ Variable	Coefficient	p value
------------	-------------	---------

◆ Mass	0.129	0.00
--------	-------	------

◆ % Ca	0.334	0.00
--------	-------	------

◆ R-square = 0.63, i.e. 63% of noise explained

# Interaction Variables

---

- ◆ 4 Explanatory Variables

◆ Mass	%Ca	Black	Black*Mass
◆ 30	5	0	0
◆ 30	5	1	30

- ◆ Black\*Mass is called an interaction variable. Here it allows one to test if the coefficient for mass differs between whites and blacks.

# Serum PSA, Ca. and Race

---

Variable	Coefficient	p value
◆ Mass	0.131	0.00
◆ % Ca	0.186	0.00
◆ Blk * Mass	0.001	0.96
◆ Blk * %Ca	0.299	0.00
◆ R-square = 0.69, i.e. 69% of noise explained		

# Serum PSA, Ca and Race

---

- ◆ Every gm of prostate contributes 0.13 ng/ml to PSA (same for blacks and whites)
- ◆ Every % of Ca contributes 0.19 ng/ml to PSA for whites
- ◆ Every % of Ca contributes an additional 0.3 ng/ml to PSA for blacks, i.e. total of  $\sim 0.48$  ng/ml.

# Remaining Noise

---

- ◆ Not reduced by adding Gleason grade
- ◆ Other variables of potential importance:
- ◆ volume of serum, amount of PSA per cell, size of vascular interface, and rate coefficients for release and degradation of PSA (Am J Clin Pathol 2002;119:80-89)

## 2 Additional Multivariate Analyses

---

- ◆ Logistic Regression Model
- ◆ Cox Model for Survival

# Probability of Disease

---

## ◆ Binary Situations:

- Sensitivity:  $P(T^+|D^+)$
- Specificity:  $P(T^0|D^0)$
- PPV:  $P(D^+|T^+)$

## ◆ Logistic:

$$P(D^+) = f(T^+, x_1, x_2, x_3, \text{etc.})$$

# Logistic Regression Model

---

- ◆ Example of a general linear model:
- ◆  $Y = \alpha + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 + \dots$
- ◆ Y can be a variety of transformations—in logistic regression Y is the logit of the probability of an outcome:
- ◆  $\text{Logit} = \log( P/(1-P) )$

# Logistic Regression Model

---

- ◆ x variables need not be binary, like the presence or absence of a positive test, but can be continuous.
- ◆ E.g. we can relate probability of prostate cancer to the exact level of PSA as well as to patient age.

# Logistic Regression Software

---

- ◆ an iterative algorithm for maximum likelihood solution
- ◆ estimates the  $\beta$  coefficients in :
- ◆  $Y = \alpha + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 + \dots$
- ◆ provides chi-square statistics to test Nulls that x variables are not important

# Logistic Model For Tests Of AMI

---

- ◆ Which test provides more information for AMI: CK-MB or Troponin I?
- ◆ Does one add information to what the other provides?

## Example: Chang et al. data from 1998

---

- ◆ 110 pts. reported.
- ◆ Sensitivities, specificities, and PPV's of CK-MB and Troponin I reported the two used separately and together:

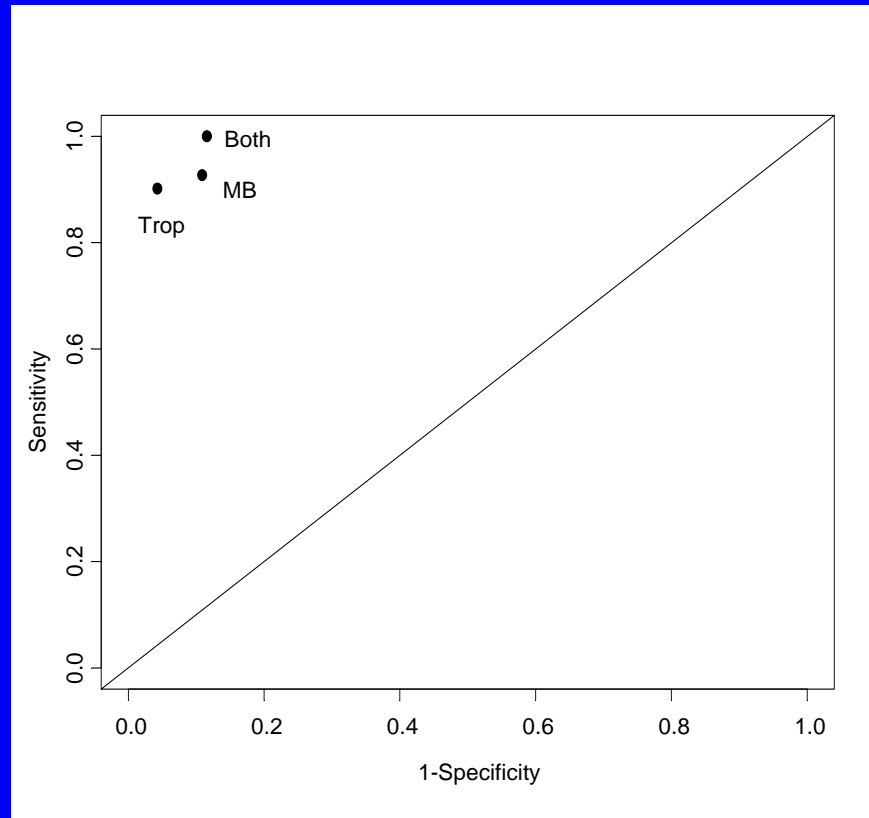
# Chang et al. Results

---

◆	Sens	Spec	PPV	NPV
◆ MB	92.7	89.9	84.4	95
◆ Trop I	90.2	95.7	92.5	94.3
◆ Both	100	88.4	83.7	100

- ◆ But what about tests of significance and how to weight these 2 tests?

# ROC for CK-MB & Trop I



# Logistic Regression of Chang Data

---

◆	Coef.	P value
◆ MB	3.47	0.00
◆ Trop I	4.18	$2.2 \times 10^{-7}$

- ◆ Together, these two variables explained 75% of deviance in data.

# Use of Logistic Coefficients

---

- ◆ Together the logistic model and the coefficients yield a diagnostic algorithm for predicting the probability of AMI.
- ◆  $E = \alpha + 3.47 * \text{CK-MB} + 4.18 * \text{Trop I}$
- ◆  $\alpha$  = the prevalence of AMI when both tests are negative.
- ◆  $P(\text{AMI}) = 1 / (1 + \exp(-E))$

# The Calculated P(AMI)

---

- ◆ Must first validate the results by comparing observed P(AMI) in a new group of pts with the calculated P(AMI).
- ◆ If they agree, then one can use this formula in new clinical settings.
- ◆ In this way the logistic model does more than test the tests--it provides an algorithm.

# Survival Analysis for a Test

---

- ◆ Sometimes, the importance of a lab test comes not so much from its association with a particular diagnosis now, but because its value relates significantly to time to an adverse outcome.
- ◆ e.g. higher levels of cholesterol imply shorter times to AMI.

# Survival Analysis

---

- ◆ Powerful and now commonly used.
- ◆ Cox model (1972) is most popular.
- ◆ 3 categories of random variables:
  - Failure Event (e.g. death)
  - Time to Event (or to censoring)
  - Explanatory  $x$  variables including the value of a lab test

# Failure Event

---

- ◆ Binary i.e. 0 or 1
- ◆ 0 means pt. censored at time of last observation.
- ◆ Could be another type of event:
  - Diagnosis
  - Tumor Recurrence
  - Metastasis
  - Response to Treatment

# Time of Analysis

---

- ◆ Usually, begins at 0 and is always positive.
- ◆ Sometimes the beginning time is missing (left censoring or truncation).
- ◆ Other, continuous positive variables are possible, but one usually assumes that the pt. has been observed continuously from zero time.

# Survival Probability, $S(t)$

---

- ◆  $S(t)$  is the probability that survival time  $T$  will exceed  $t$ , i.e.
- ◆  $S(t) = P(T > t)$
- ◆  $S(0) = 1.0$  and  $S(\infty) = 0$ .
- ◆ Kaplan-Meier plot: graph of  $S(t)$  vs.  $t$ .

# Non-Small Cell Lung Ca (NSCLC)

---

- ◆ ~30% Localized or Resectable
  - Rx: Surgery
- ◆ ~70% Advanced Stage and Non-Resectable
  - Rx: RadRx, Chemotherapy
  - Pilot Study of Serum CYFRA 21-1 as early measure of response

# Serum CYFRA 21-1

---

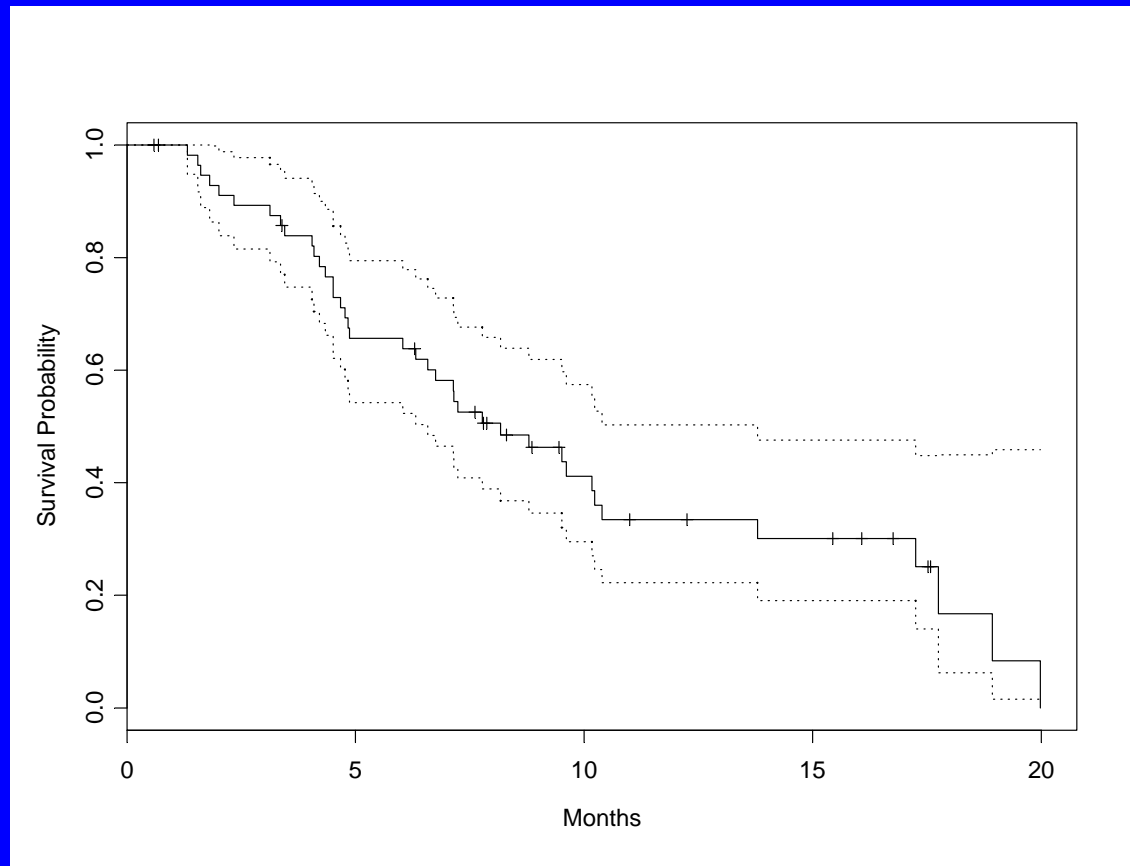
- ◆ Electrochemiluminescence immunoassay for soluble cytokeratin 19 protein fragments in the serum.
- ◆ Not helpful for screening in NSCLC.
- ◆ However, serum CYFRA
  - correlates with stage
  - correlates with survival
  - decreases after effective surgical excision

# Example of Survival Data

---

◆ Stage	Dead	T (months)
◆ IIIb	1	18.9
◆ IV	1	3.1
◆ IV	0	12.2

# Kaplan-Meier Plot in Pilot Study of Advanced NSCLC, n=58

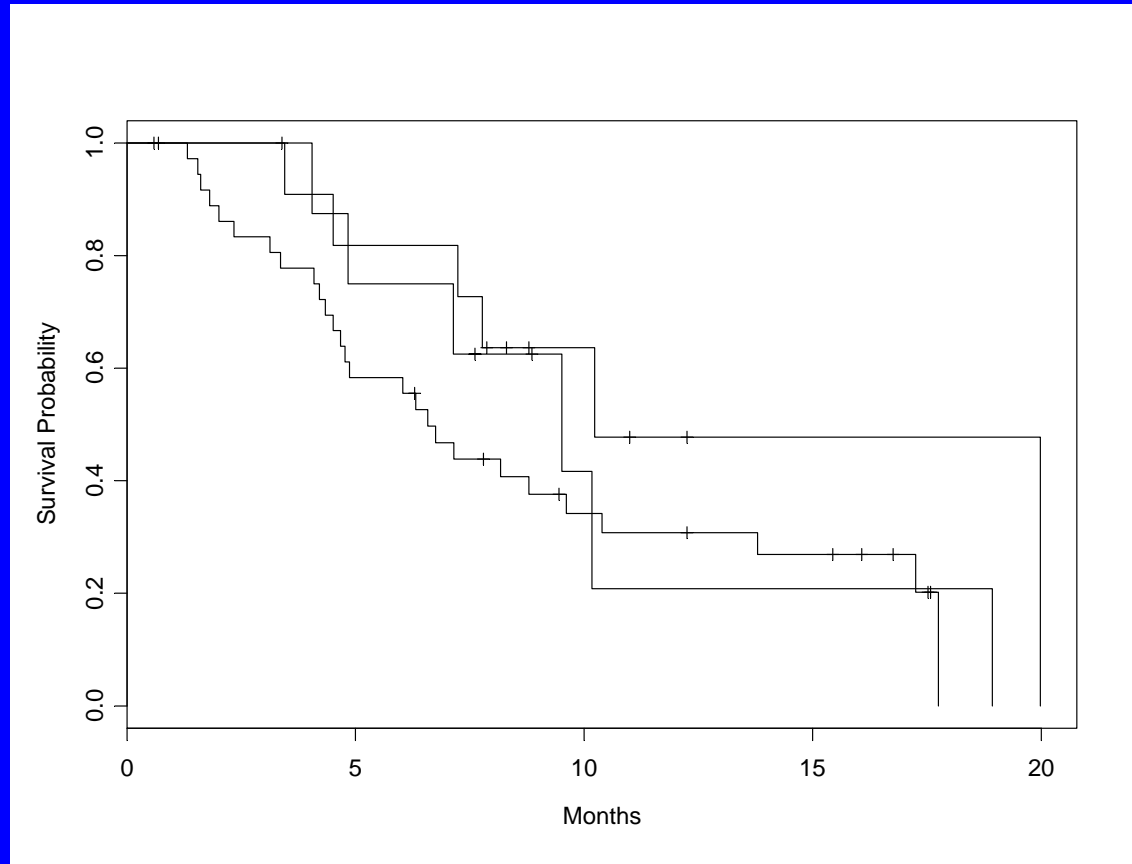


# Motivation of Survival Analysis

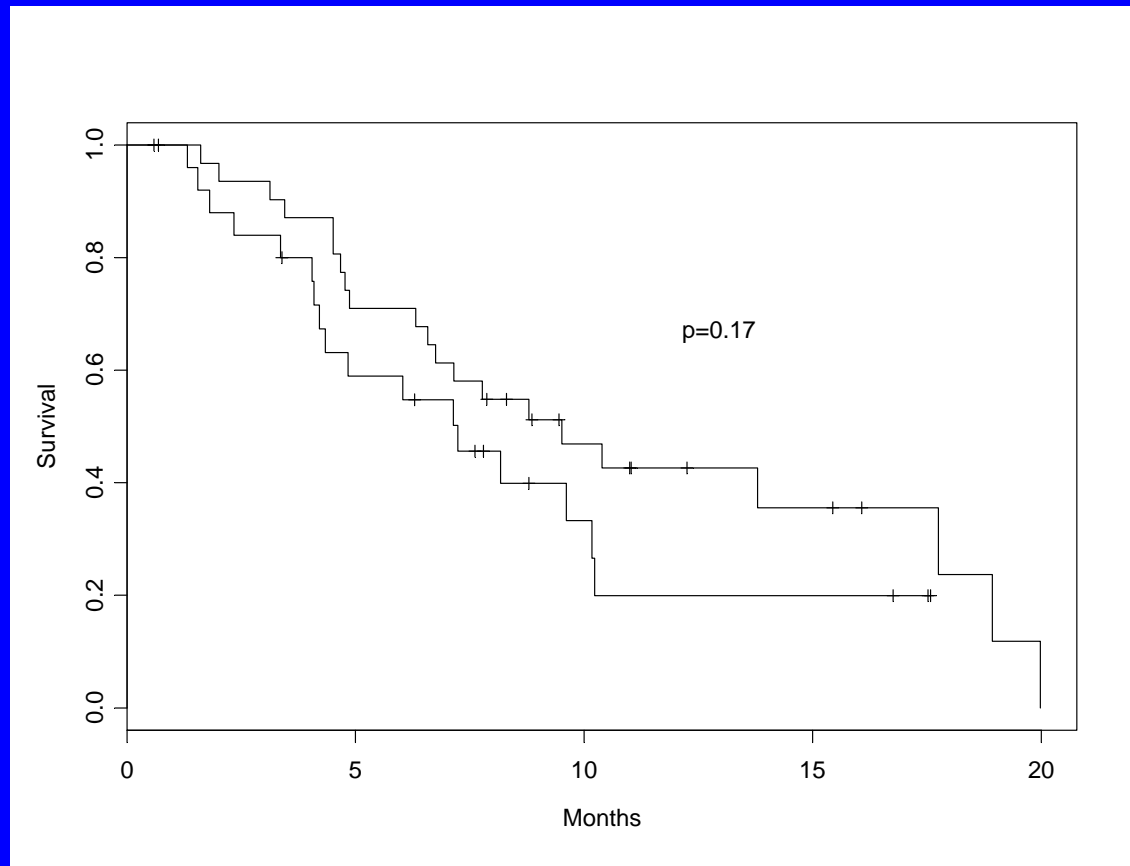
---

- ◆ To test whether certain variables affect survival time.
- ◆ Univariate-graphical: Kaplan-Meier Plots
- ◆ Univariate-statistical: log-rank test
- ◆ Multivariate-Cox Model

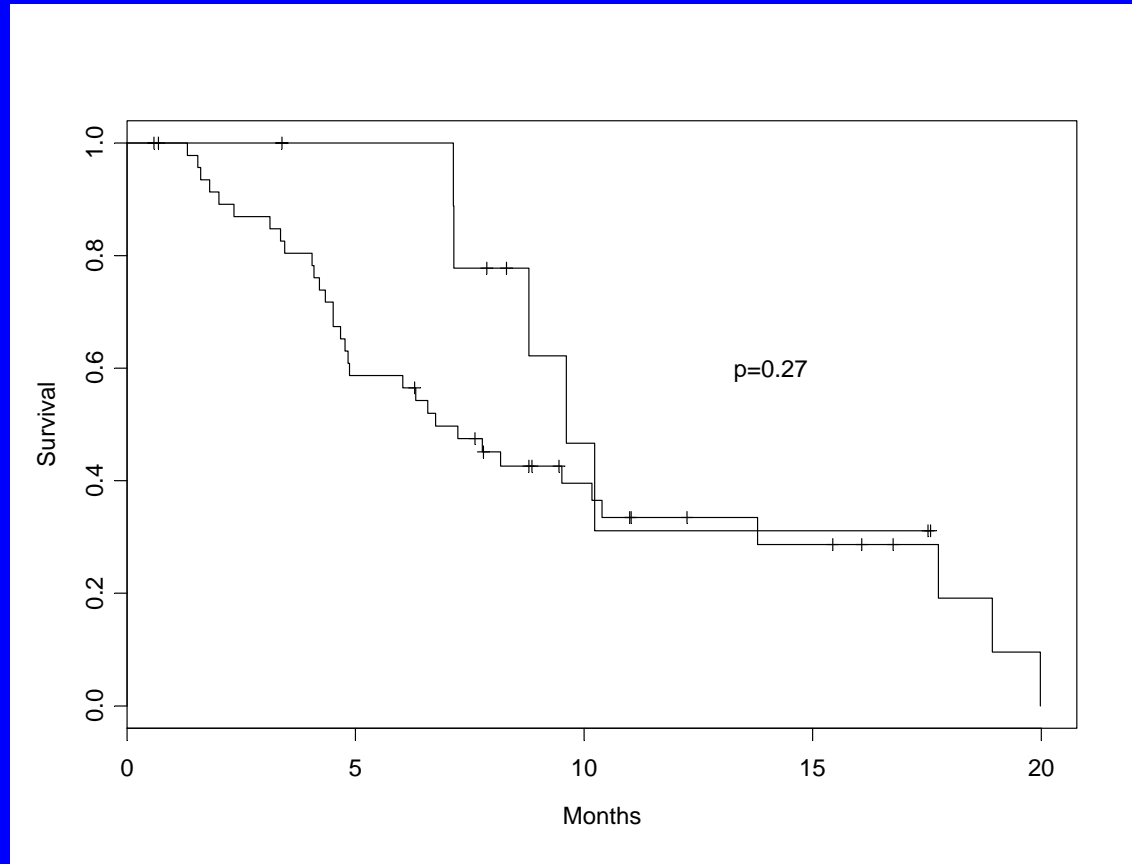
# Survival by Clinical Stage for Cyfra Study



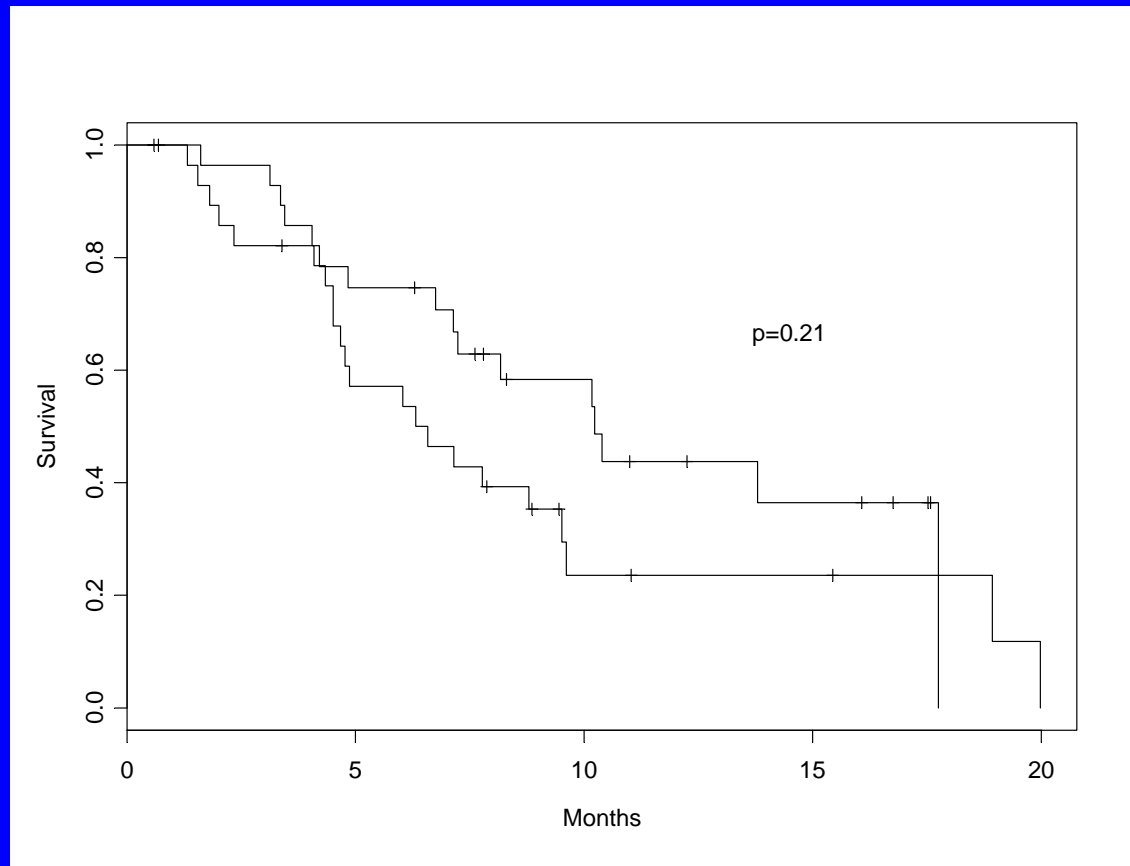
# Survival vs. Initial CYFRA ( $\leq$ vs. $>$ 3.9 ng/ml)



# Survival vs. Response (PR or Not)



# Survival vs. Cyfra Drop > 27%



# Univariate Test--Log Rank Test

---

- ◆ Compares observed deaths in a subgroup with those expected to occur from whole data without subgroups.
- ◆ Results in a Chi-square statistic.

# Log-Rank for Cyfra Study

---

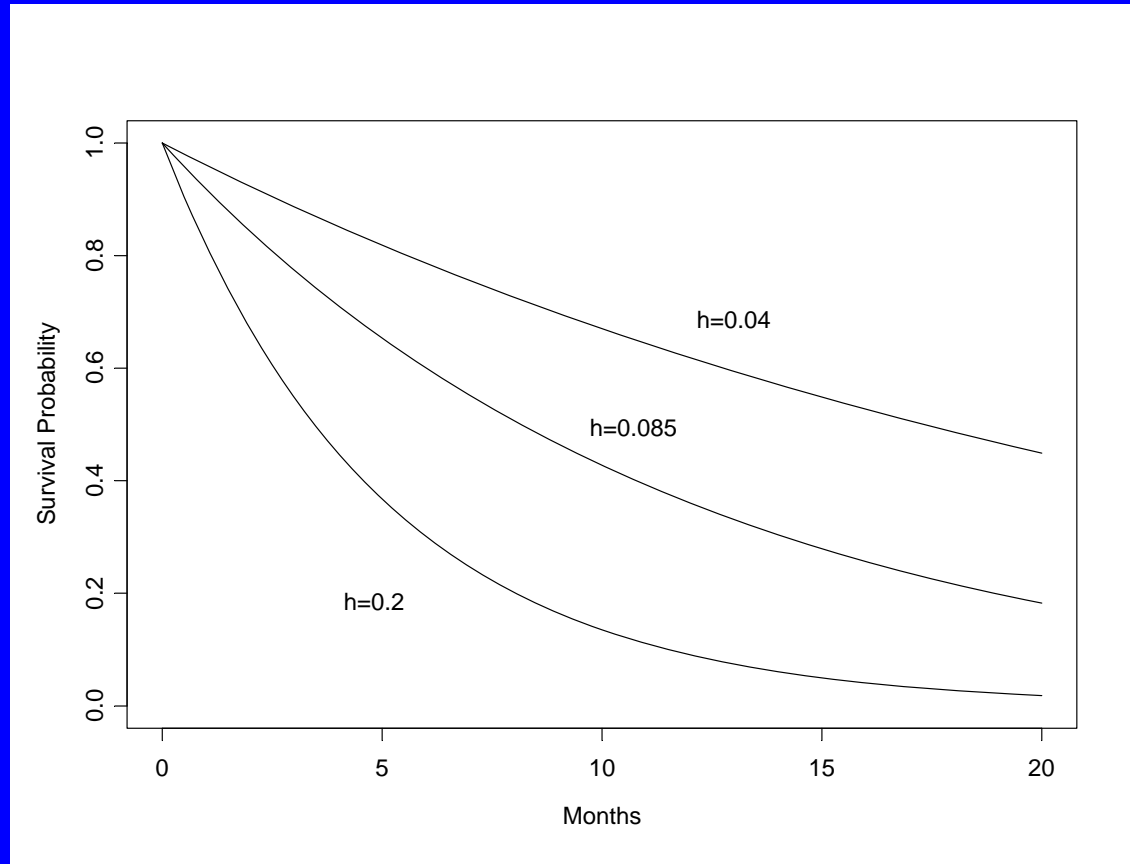
◆ Variable	p value
◆ Stage	> 0.2
◆ Cyfra	0.17
◆ PR	0.27
◆ Drop in Cyfra	0.21

# Concept of Hazard

---

- ◆ Shapes of most survival curves suggest an exponential type function:
- ◆  $S(t) = \exp(-h \cdot t)$
- ◆  $h$  is the hazard.
- ◆ The larger the value of  $h$ , the steeper is the slope of the curve and the shorter is the survival.

# Concept of Hazard



# Calculus of $S(t)$ and Hazard, $h$

---

- ◆  $S(t) = \exp \left\{ - \int_0^t h(u) du \right\}$  with integration from 0 to  $t$  or equally from the derivative:
- ◆  $d \ln( S(t) )/dt = - h (t)$
- ◆ Thus,  $h(t)$  is the instantaneous slope of the curve of  $\ln( S(t) )$  vs.  $t$ .

# Cox Proportional Hazards Model

---

- ◆ The Cox model evaluates the relationship between the hazard,  $h$ , and possible explanatory variables.

# The Cox Model

---

- ◆ A proportional hazards model.
- ◆ If  $h_0(t)$  is an unspecified baseline hazard function, then the Cox model assumes that
- ◆  $\ln(h/h_0) = \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 + \dots$
- ◆  $(x_1, x_2, x_3, \dots)$  are the explanatory variables.

# The Cox Model

---

- ◆ Software obtains:
  - maximum partial likelihood solutions for the  $\beta$  coefficients and
  - chi-square tests for the Null that the x variables are not related to survival time.

# The Cox Model Assumptions

---

- ◆ the ratio  $h/h_0$  should not vary with time.
- ◆ S-PLUS software provides tests for the suitability of this assumption.

# Power and Sample Size

---

- ◆ General rule: 10 uncensored pts. for every explanatory variable used.
- ◆ e.g. Cyfra study had 39 of 58 uncensored.

# Cox Model for Cyfra Study

---

◆ Variable	p value
◆ Stage IIIa	0.14
◆ Stage IIIb	0.16
◆ PR	0.18
◆ Log(Initial CYFRA)	0.000028
◆ Cyfra Drop > 27%	0.0035

# Conclusions for CYFRA Study

---

- ◆ The level of CYFRA before treatment is more important than exact clinical stage in advanced NSCLC.
- ◆ The drop in CYFRA after the first cycle of Chemo Rx is more important than image based response.

# Hazard Score

---

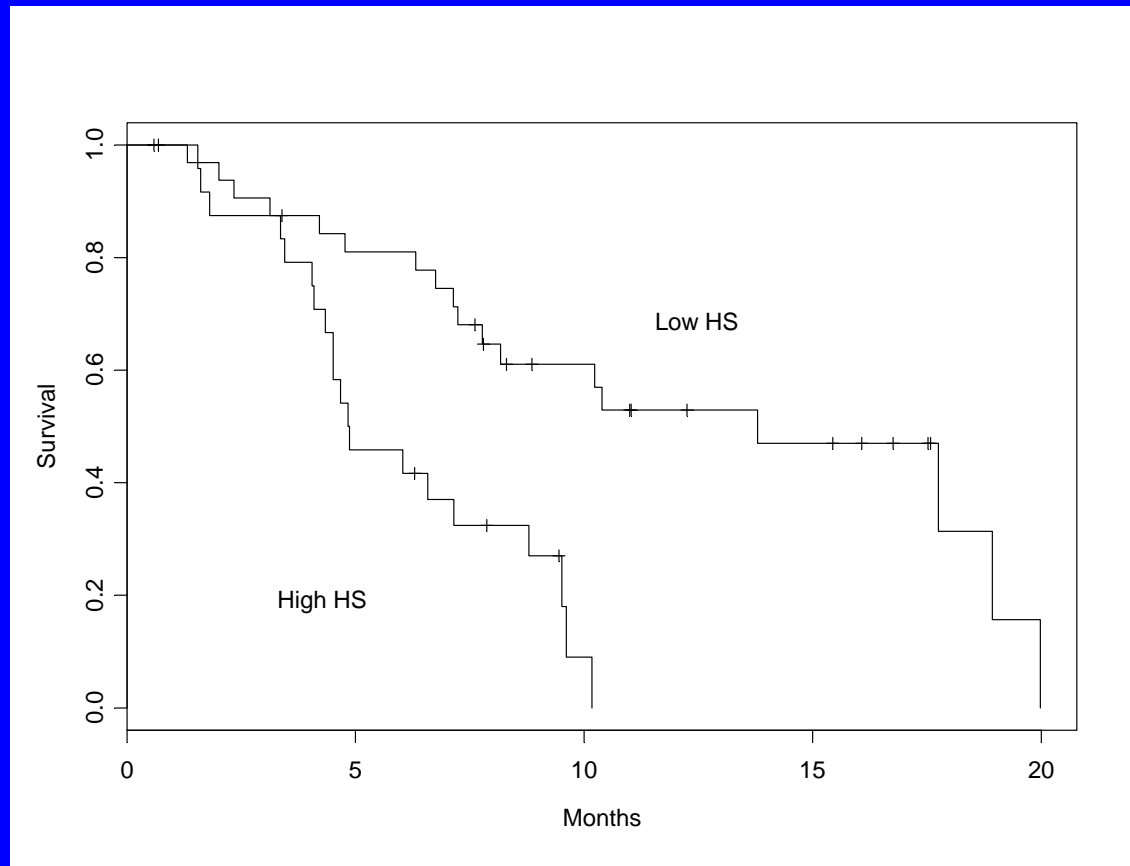
- ◆ To consolidate the prognostic information of all the variables use a hazard score (HS).
- ◆ HS is just the sum of the Cox model coefficients times their variables.

# Hazard Score For CYFRA

---

- ◆  $HS = 0.61 * \log(\text{Initial CYFRA}) - 1.14 * R$
- ◆ (R is 1 if there was > 27% drop in CYFRA after 1st cycle of chemoRx. Otherwise, R is 0.)
- ◆ Here, HS uses results of 2 serum samples: one before and one after the 1st Rx.

# Effect of Combined Hazard Score (HS) on Survival



# Hypothesis From Cyfra Study

---

- ◆ In a disease where the best we can expect is 20-30% response to Rx, the combined HS after the 1st cycle of Chemo Rx could help decide whether to:
  - ◆ 1) continue same Rx
  - ◆ 2) switch Rx
  - ◆ 3) stop Rx and decrease suffering and costs

# Portability of Cox Models

---

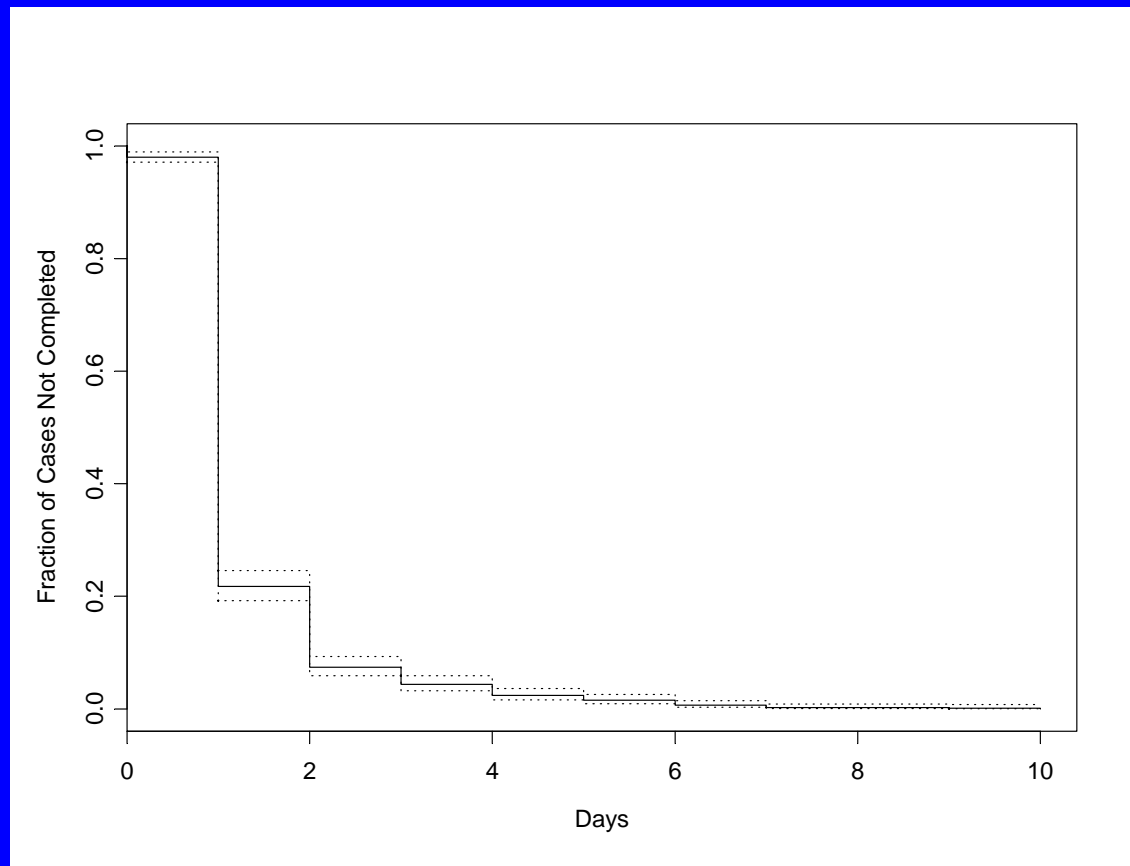
- ◆ If the survival analysis has been validated with new data, then the Cox model and its coefficients can be used to predict survival for new patients.

# Laboratory TAT's

---

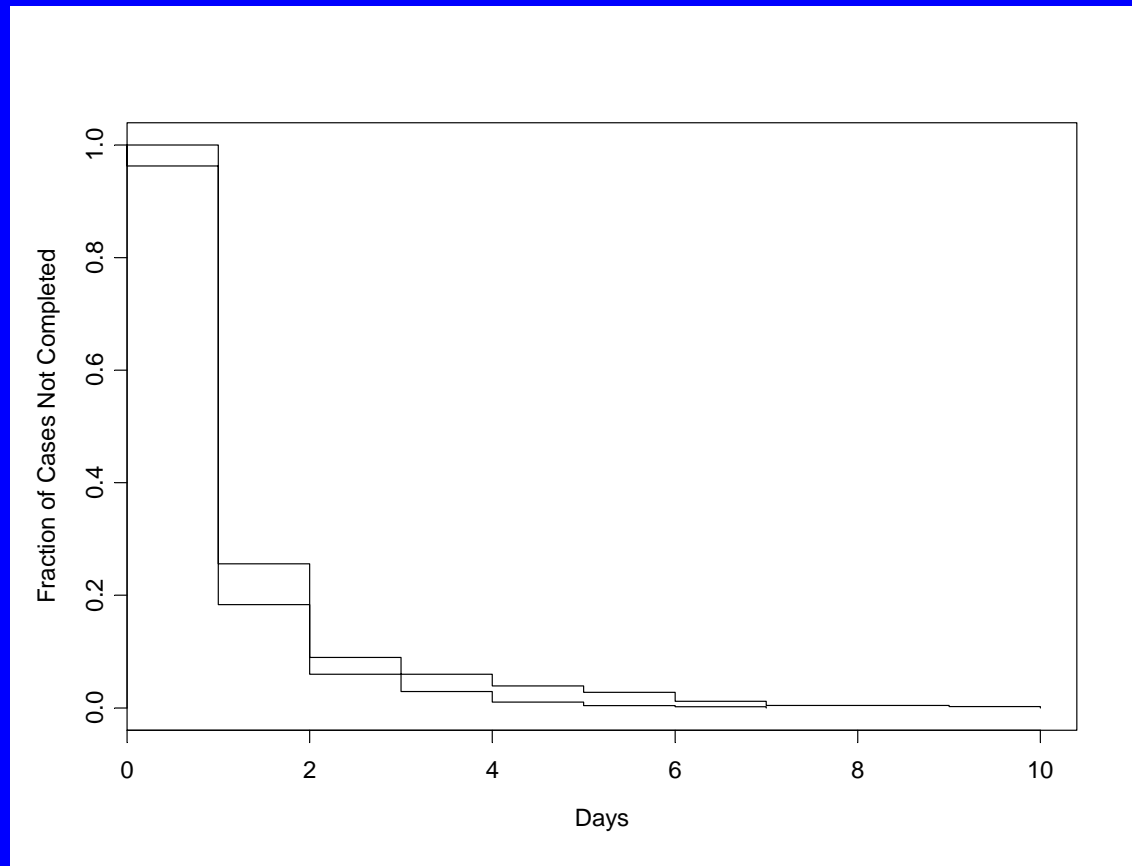
- ◆ Like survival times, laboratory TAT's are always greater than or equal to 0
- ◆ The reporting of a test result is its “death” event in the lab.
- ◆ All of the foregoing techniques of survival analysis can be used to study TAT.

# Kaplan-Meier Plot of TAT in Surgical Pathology at Durham VAMC



# TAT for Two Pathologists

---



# TAT for Two Pathologists

---

- ◆ Pathologist A: 1.25 days
- ◆ Pathologist B: 1.5 days
- ◆ Log rank test: P value  $\sim 0$
- ◆ Clinical significance: Nil
- ◆ Other key variables?

# Lymph Nodes in Colon Ca

---

- ◆ Pts with pos. nodes do worse and benefit from chemoRx.
- ◆ Empirically, the chance of finding pos. nodes increases with the number of examined nodes.

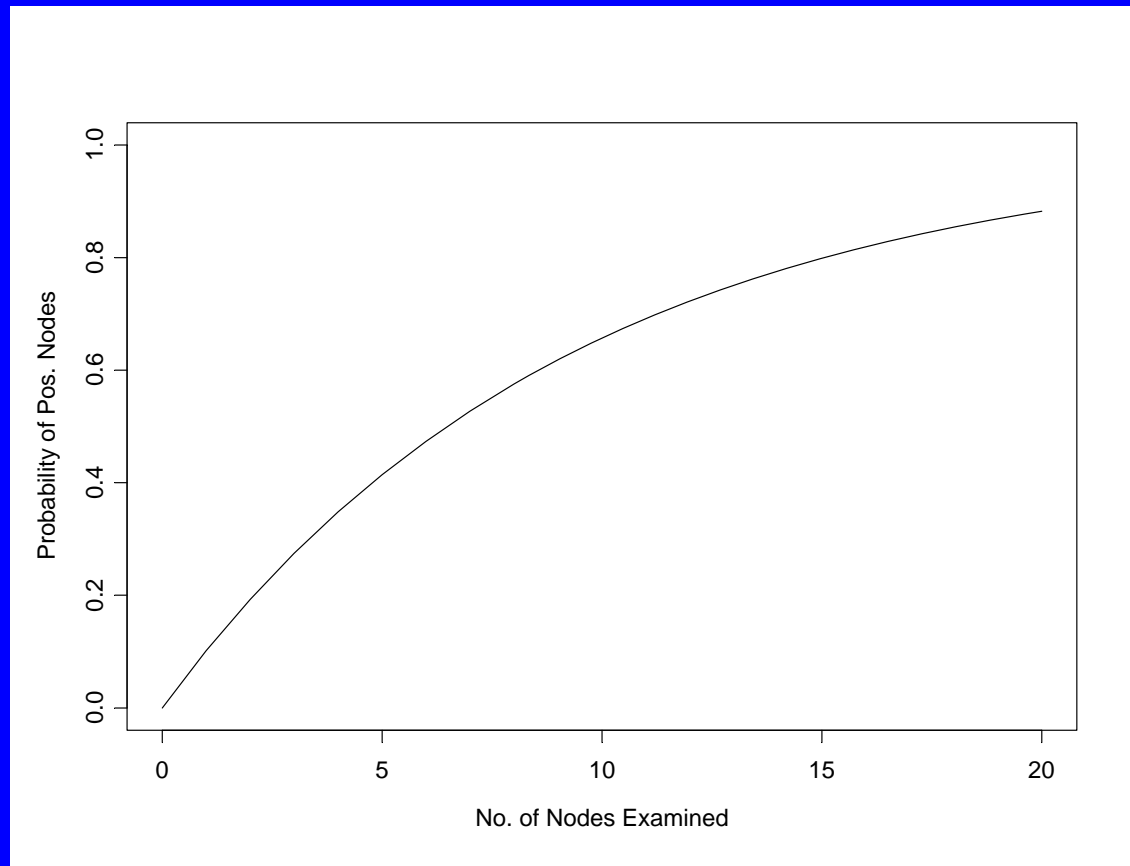
# Poisson Probability Paradigm

---

- ◆ The Poisson probability distribution function deals with the number of positive nodes observed,  $x$ , when  $n$  total are found.

# Poisson e.g. for T2 Tumor

---



# Poisson Probability Paradigm

---

◆  $(\alpha * n)^x * \exp(-\alpha * n)$

◆  $P(x) = \frac{\text{-----}}{\text{-----}}$

◆  $x!$

◆  $x$  = no. pos. nodes,  $n$  = total nodes,  $\alpha$  is the underlying per node chance of metastasis.

# Poisson Probability Paradigm

---

- ◆  $\alpha$  is the underlying probability of the patient having positive lymph nodes.
- ◆ Two possibilities exist:
  - ◆ 1) If  $\alpha = 0$ , then the patient has no metastases.
  - ◆ 2) If  $\alpha > 0$ , then the patient has metastases.

## Estimating $\alpha$ when metastases are present

---

◆ In 157 pts with positive lymph nodes,  $\alpha$  was significantly related to T stage.

◆ T Stage	$\alpha$
◆ T1	0.091
◆ T2	0.125
◆ T3&4	0.304

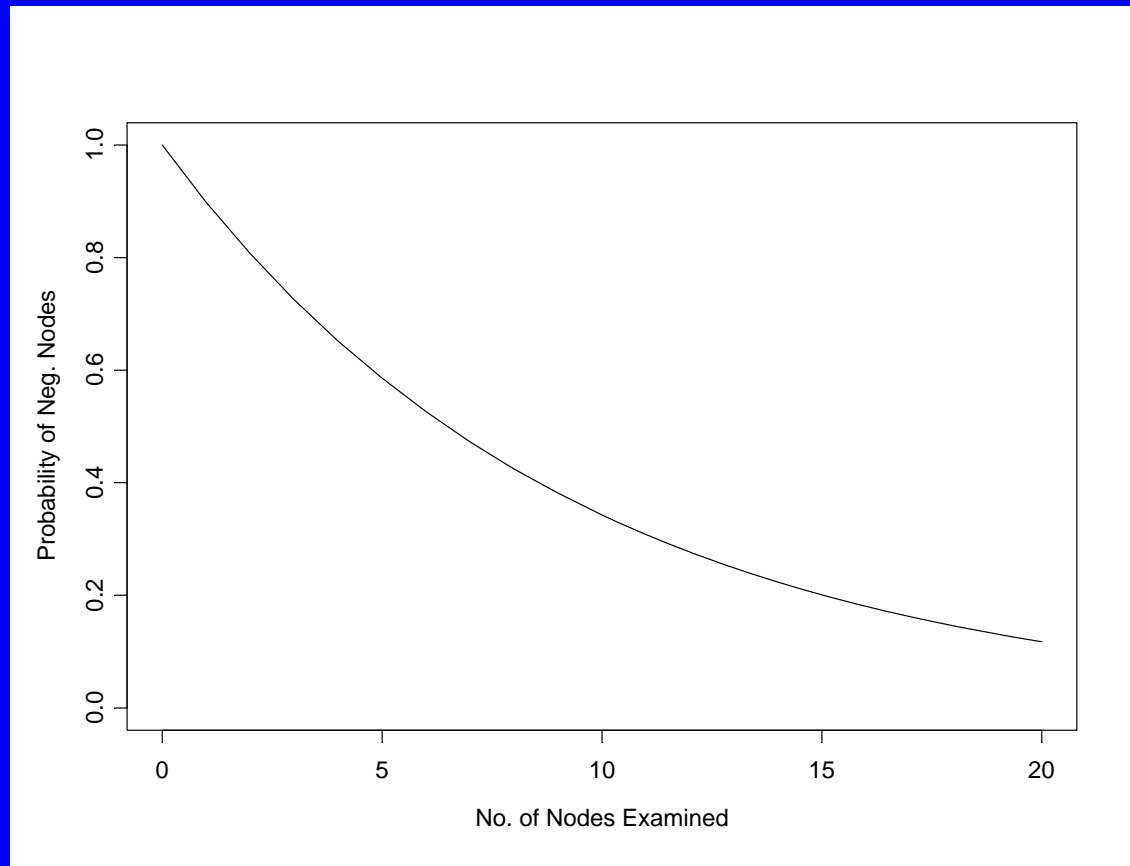
# Poisson Paradigm

---

- ◆ For pts. with metastases (i.e.  $\alpha > 0$ ) there is a real chance of missing metastases when the no. of examined nodes is small.

# Probability of finding negative lymph nodes for T2 Tumor with $\alpha > 0$

---



# Bayes' Theorem

---

$$\begin{aligned} & \diamond \quad P(T+|D+) * P(D+) \\ \diamond \quad P(D+|T+) &= \frac{P(T+|D+) * P(D+)}{P(T+|D+) * P(D+) +} \\ & \diamond \quad P(T+|D0) * (1-P(D+)) \end{aligned}$$



# With Poisson Model

---

- ◆  $P(x=0|\alpha>0) = \exp(-\alpha * n)$
- ◆  $P(x=0|\alpha=0) = 1.0$
- ◆ P(Mets) estimated from prior data in the literature, etc.

# Bayes' Theorem

---

- ◆
- ◆  $P(\text{Mets} | x=0) = \frac{1}{1 + \frac{\text{Odds of No Mets}}{\exp(-\alpha * n)}}$
- ◆
- ◆
- ◆

# 213 Pts with N0 Colon Ca

---

- ◆ Median no. of nodes = 8 (range 1 to 69)
- ◆ All nodes were neg. for CA.
- ◆ T Stage: T1 12%, T2 32%, T3 56%
- ◆ Bayes  $P(\text{Met} \mid x=0)$  was calculated:
- ◆ Mean = 0.11, Range: ~0.0 - 0.45

# Cox Survival Analysis

---

◆ Variable	p value
◆ Bayes Probability	0.0006
◆ T Stage	> 0.8

# Survival Plot for N0 Colon Ca

