Session II

Evaluating Discriminatory Power and Forecast Performance

C. Erik Larson

Lead Expert for Enterprise Risk
Risk Analysis Division
Washington, DC 20219
Erik.Larson@occ.treas.gov

Outline

- Modeling Objectives
- Traditional Credit Risk Model Design
- Discriminatory Power
- Forecast Performance
  - Business Decisions
  - Accuracy and Precision
  - Evaluating Rating or Score Level Prediction
  - Evaluating Global Fit

Modeling Objectives

- Should be linked to business needs and use
- Can influence:
  - the logical design of the model
  - the sampling design
  - the statistical techniques employed in estimation
  - the benchmarking and performance tracking techniques
  - the interpretation of validation results
- Should generally be determined early in the modeling process

Modeling Objectives

Discrimination and Prediction

- The qualitative or ordinal **discrimination** between two or more types of credit
- *Examples:*
  - *Risk ranking of delinquent borrowers to allocate followup-efforts*
  - *Segmentation of applications for different review*
- The **forecasting** of cardinal risk levels for individual credits
- *Examples:*
  - *Default probability estimation*
  - *Loss forecasting*

Traditional Credit Risk Model Design

- Default, delinquency and segmentation models have traditionally been developed to meet a classification objective.
- The dependent variable of interest takes a <u>limited set of values,</u> {0,1}, corresponding to membership in a class.
- *Examples:*
  - *Good vs. Bad*
  - *Non-Delinquent vs. Delinquent*
  - *Non-Default vs. Default*
  - *Low Risk vs. High Risk*

Traditional Credit Risk Model Design

- Rating and scoring models develop predictions of class membership as a function of borrower characteristics, X_i.

- Typical Model
  - The score: $z_i = Z(X_i, \hat{\beta})$
  - Implementation:
    Choose a score cutoff z^*.
    If borrower's score is less than cutoff, predict bad;
    if score is greater than or equal to cutoff, predict good.

- Let $F(z|Good)$ and $F(z|Bad)$, respectively, represent the cumulative distribution functions of "good" borrowers and "bad" borrowers generated by the score.

- Question: How should z^* be chosen?

Discriminatory Power

Choosing a Score Threshold: Types of Errors

- Model predictions of class membership are compared to realized outcomes

		Realized Outcome	
		Good	Bad
Predicted Outcome	Good $z_i > z^*$	No Error	Type I Error
	Bad $z_i <= z^*$	Type II Error	No Error

- This is closely related to retail scorecard "swap-set" analysis

Discriminatory Power
One Score Threshold: The K-S Statistic

- One way to choose z^* is by picking a value that minimizes expected costs from making Type I and Type II errors:

$$c_b \text{ Prob[Type I Error]} + c_g \text{ Prob[Type II Error]}$$

$$c_b \,(1-F(z^*|\text{Bad}))\text{ Prob[Bad]} + c_g \, F(z^*|\text{Good})\text{ Prob[Good]}$$

- Here "c_b" is the cost from making a loan that turned out bad, and "c_g" is the opportunity cost of failing to make a loan that would have turned out good.

- Note that if c_gProb[Good] = c_bProb[Bad], then this problem reduces to maximizing

$$F(z^*|\text{Bad}) - F(z^*|\text{Good})$$

- *This is equivalent to setting the score threshold at the value for which the K-S statistic is maximized (see Thomas, et. al. [2002])*

A problem with this argument....

- We seldom observe approve/decline decisions made by setting a cutoff equal to the score value which maximizes K-S.

- In fact, we usually see many thresholds used in decisioning. What should we conclude?

- ***The use of K-S to evaluate a model's discriminatory power might not provide insight into the model's performance in the range required by the business decision*** *(Hand [2004])*

- ***Other metrics might be needed.***

Forecast Evaluation
Accuracy and Precision

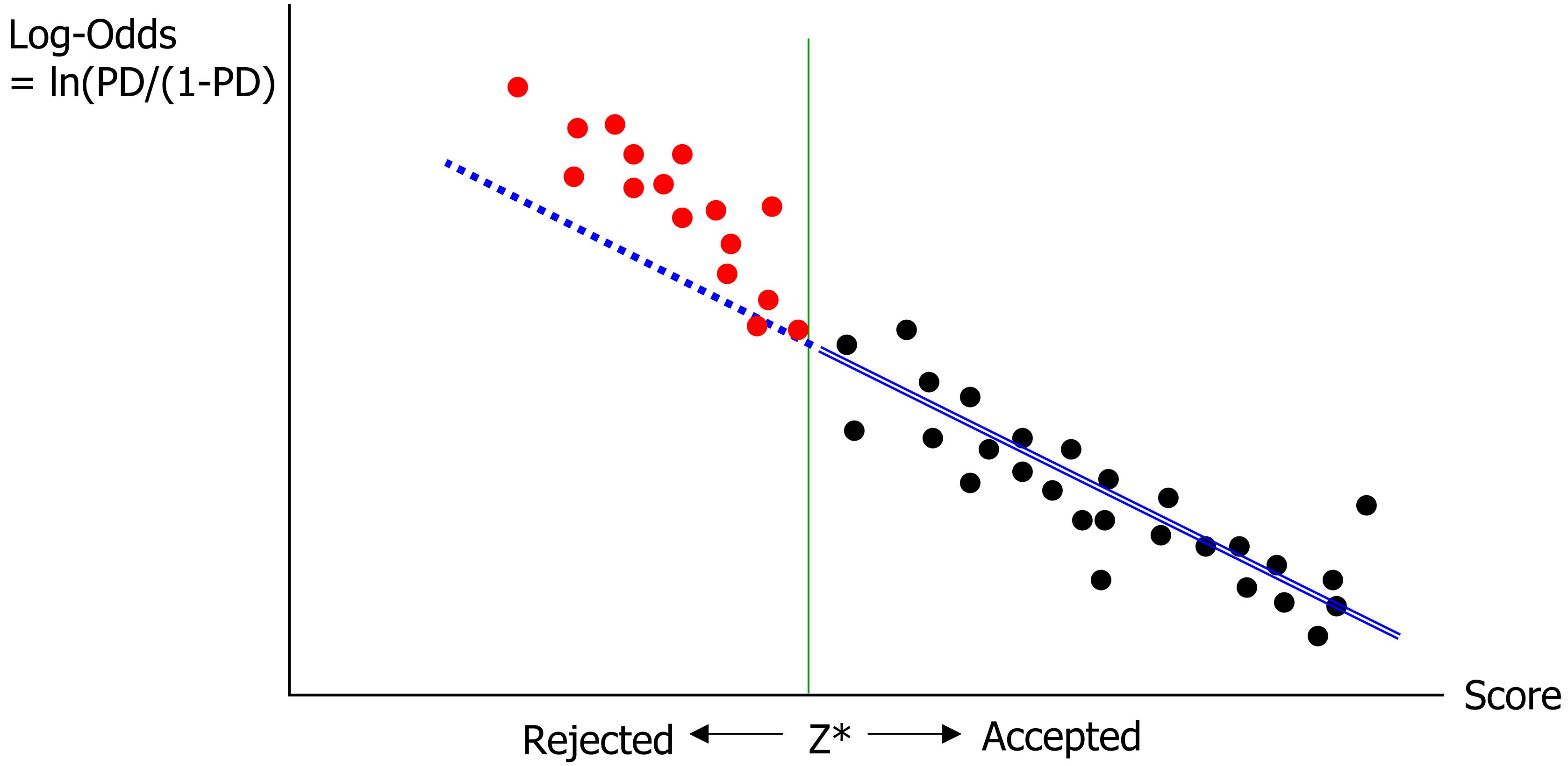
- The concepts of **accuracy** and **precision** can be employed when evaluating rating and scoring model performance at a number of different thresholds.

- A forecast is considered *accurate* if it is "right" on average, i.e. if the predicted outcome on average coincides with the actual outcome. This concept of accuracy is closely related to the *unbiasedness* of a statistical estimator.

- *Precision* is usually defined as the inverse of the standard error (or variance) of an estimator. Less precision is reflected by a larger standard error.

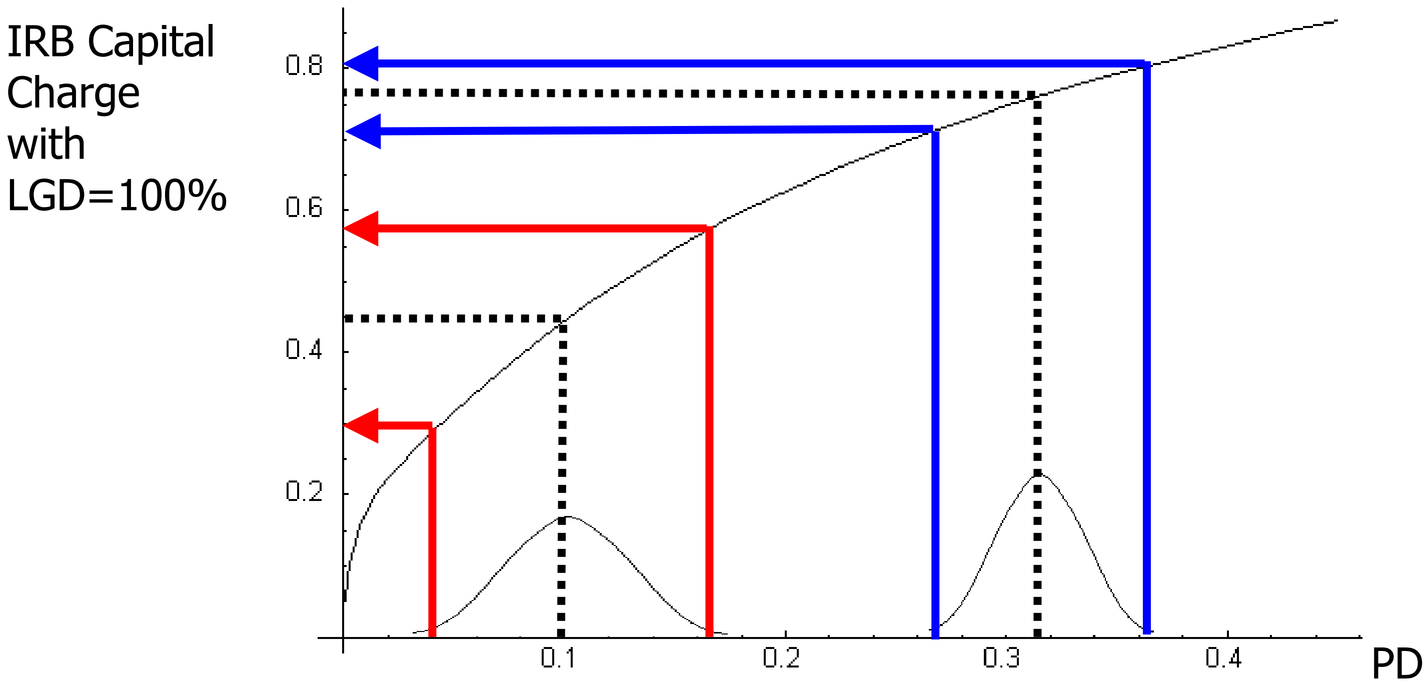
Business Decisions Influenced by Forecast Accuracy

- Reject-inference (prediction of performance for rarely-booked credits)



Business Decisions Influenced by Forecast Precision

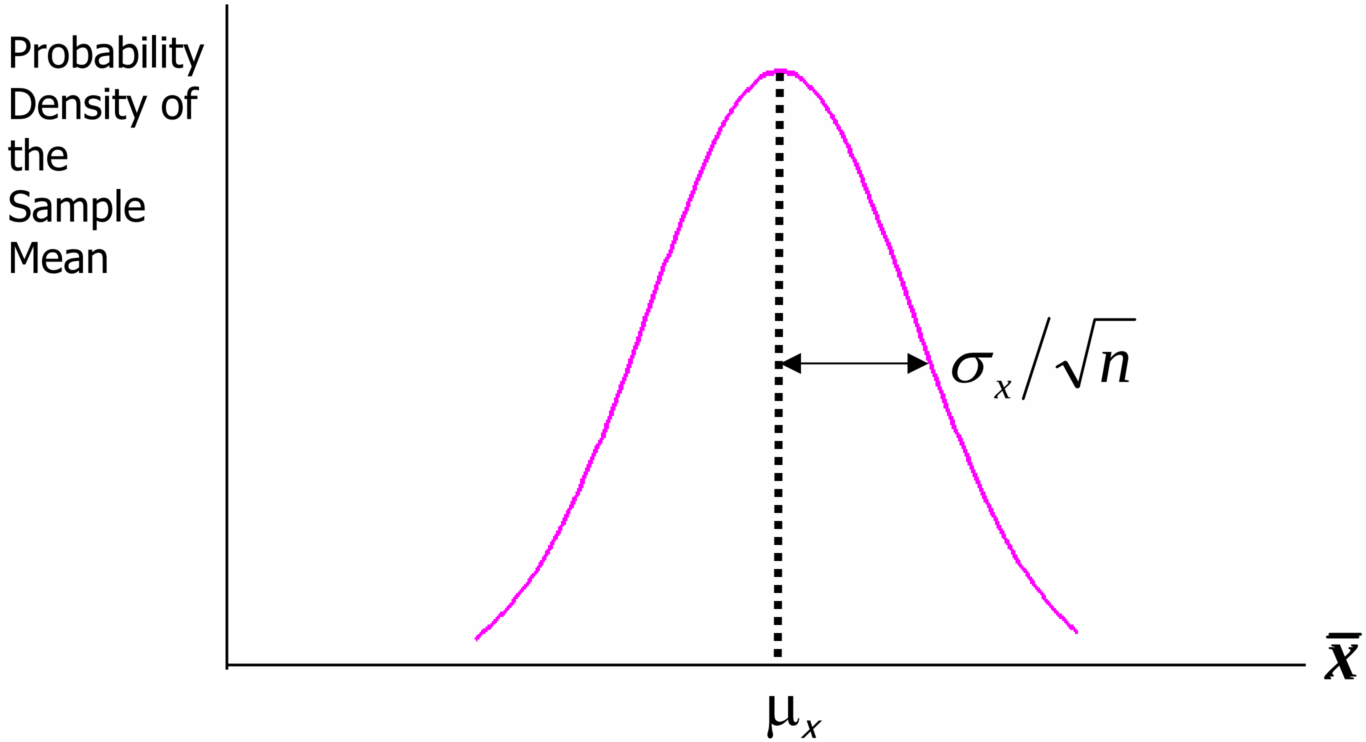
- The variability in capital that could be induced by variability in PD.



Source: Basel II formula for corporates, sovereigns, and banks(BCBS [2005]).

Statistical Evaluation of Accuracy and Precision

- The **Central Limit Theorem** tell us that that when based upon a sufficiently large sample, the sample mean of an estimator, $(\bar{x})$, will be distributed normally around the true population mean (μ_x), with a standard deviation equal to the population standard deviation (σ_x) divided by the square root of the sample size (n).



Evaluating Rating or Score Level Forecasts
The Interval Test

- To examine the accuracy and precision of a PD or LGD forecast for an individual rating grade, we can use the Central Limit Theorem to construct a test of the null hypothesis that the true mean is equal to the predicted value for the grade. We then compare the observed value of PD or LGD with this interval.

- We construct a 95% confidence interval as

  Parameter Estimate +/- 1.96*Parameter Standard Error

- When focusing on PD, the standard error can be computed as SquareRoot(PD*(1-PD)/N), where N equals the number of observations in a rating bucket. The interval is computed as ranging from

$$PD-1.96\times\sqrt{\frac{PD\times(1-PD)}{N}} \quad \text{to} \quad PD+1.96\times\sqrt{\frac{PD\times(1-PD)}{N}}$$

Example

Rated loan portfolio for RMH Bank

RMorrisHamilton

Grade Default Report, December 31, 2002

Period covered by report: 1997 – 2002

1 Internal Grade	2 Estimated PD (12-31-02)	3 Number of Obligors	4 Portfolio Share	5 Number of Defaults	6 Actual Default Rate
1	0.03%	3660	4.5%	3	0.08%
2	0.05%	5800	7.2%	5	0.09%
3	0.25%	9500	11.8%	10	0.11%
4	1.20%	38200	47.5%	217	0.57%
5	5.50%	21240	26.4%	396	1.86%
6	11.00%	1100	1.4%	111	10.09%
7	15.00%	990	1.2%	177	17.88%
Total		80490	100%	919	1.14%

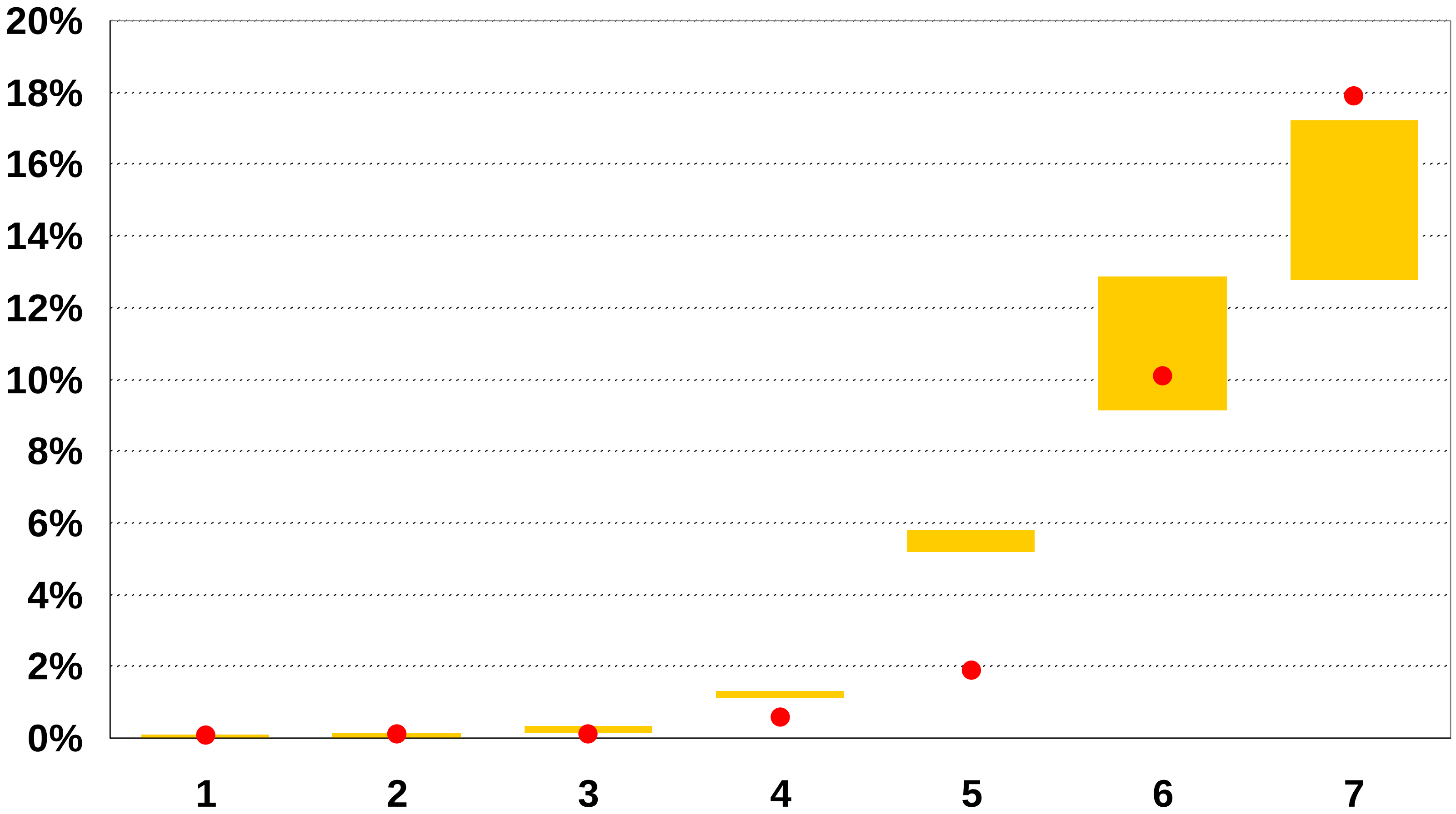
Example

Interval Tests for RMH Bank's PD Estimates

Rating Grade	Expected Default Rate (PD)	N	Standard Error	Confidence Interval Lower	Upper	Actual Default Rate
1	0.0003	3660	0.000286	0.000	0.001	0.0008
2	0.0005	5800	0.000294	0.000	0.001	0.0009
3	0.0025	9500	0.000512	0.001	0.004	0.0011
4	0.0120	38200	0.000557	0.011	0.013	0.0057
5	0.0550	21240	0.001564	0.052	0.058	0.0186
6	0.1100	1100	0.009434	0.092	0.128	0.1009
7	0.1500	990	0.011348	0.128	0.172	0.1788

$$PD - 1.96 \times \sqrt{\frac{PD \times (1 - PD)}{N}} \quad \text{to} \quad PD + 1.96 \times \sqrt{\frac{PD \times (1 - PD)}{N}}$$

Example
Interval tests for RMH Bank's PD Estimates



(Bars denote 95% confidence interval around grade PD; dots are actual realized default rates for each grade.)

Evaluating Forecast Performance Globally

The Chi-Square Test

- The Chi-Square Goodness-of-Fit statistic (Pearson [1900]) can be used to test the null hypothesis that the observed data follow a specified distribution.

- If there are k grades and c=2 states (default and non-default) then we are testing a null hypothesis about k binomial random variables. If the outcomes for each grade are independent, then the joint test will be distributed as a Chi-Square random variable with k degrees of freedom.

- The observed (O) and expected (E) frequencies of default and non-default are compared for each grade, and the statistic is computed as:

$$\chi^2 = \sum_{i=1}^{kc} (O_i - E_i)^2 / E_i$$

Example

The Chi-Square Test for RMH Bank's PD Estimates

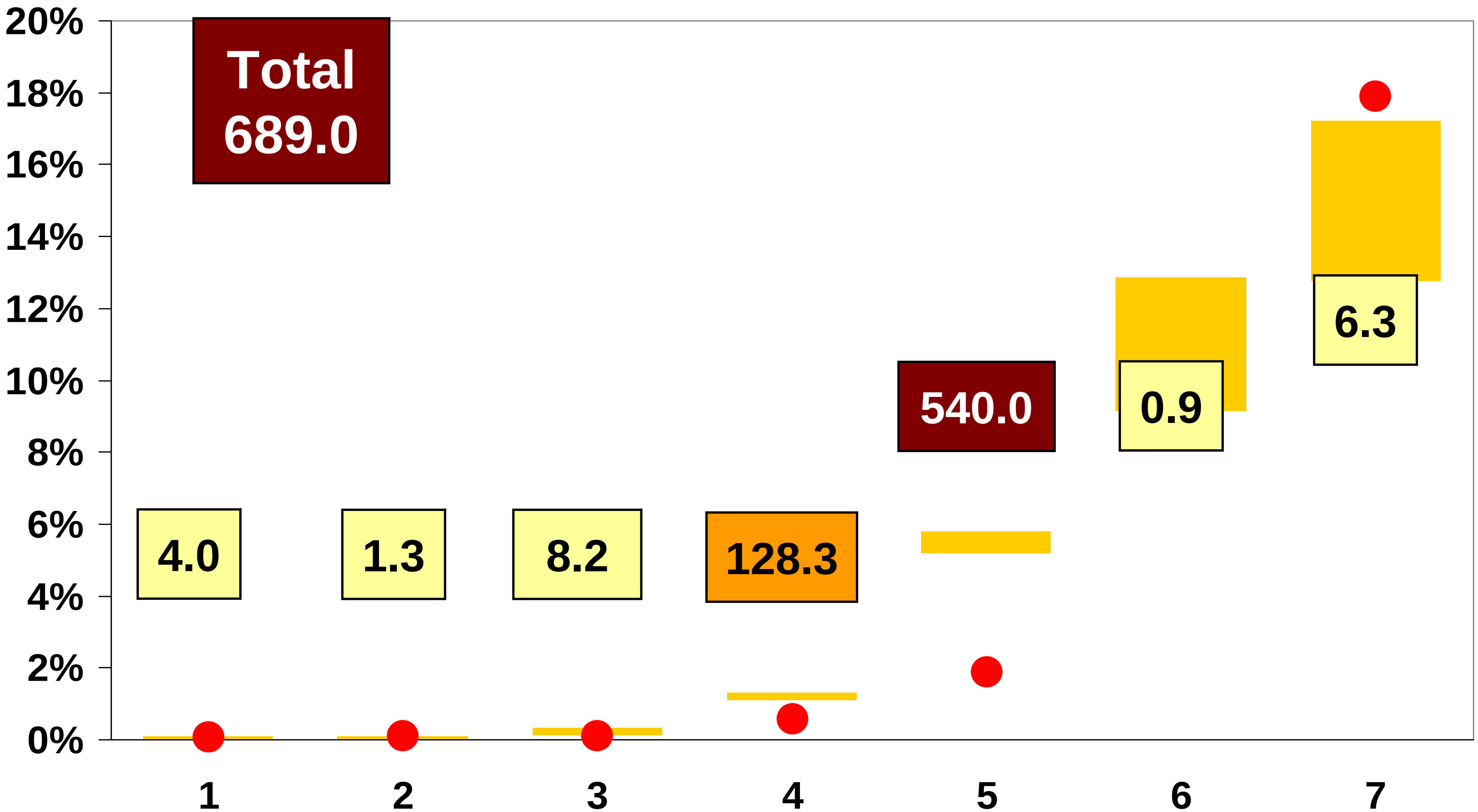
Rating Grade	PD	N	Observed Default	Observed Non-Default	Expected Default	Expected Non-Default	ChiSq Contrib Default	ChiSq Contrib Non-Default
1	0.0003	3660	3	3657	1	3659	4.00	0.00
2	0.0005	5800	5	5795	3	5797	1.33	0.00
3	0.0025	9500	10	9490	24	9476	8.17	0.02
4	0.012	38200	217	37983	458	37742	126.81	1.54
5	0.055	21240	396	20844	1168	20072	510.26	29.69
6	0.11	1100	111	989	121	979	0.83	0.10
7	0.15	990	177	813	149	842	5.26	1.00

Chi-Square Statistic=	689.02
Degrees of Freedom =	7
Prob(Chi-Sq>Critical Val)=	0.00

$$\chi^2 = \sum_{i=1}^{kc} (O_i - E_i)^2 / E_i$$

Example

The Chi-Square Test for RMH Bank's PD Estimates



(Bars denote 95% confidence interval around grade PD; dots are actual realized default rates for each grade.)

Be careful with these tests!

- Default rates are very low for most grades. With such low default rates, need a very large number of loans to achieve desirable levels of statistical confidence.

- The tests assume that defaults in each grade are independent, and they almost certainly are not.

- The tests assume that the “true” default rate is constant, and it almost certainly is not.

- ***The practical implication is that the true 95% confidence bands for the PD estimates are probably wider than derived.***

Testing Global Accuracy
Other Related Tests

The Chi-Squared test's sensitivity to how observations are distributed across the k grades has led to the development of some alternative tests:

- The **Hosmer-Lemeshow Test** (Hosmer and Lemeshow [2000])

  A Chi-Square test where the data is regrouped into deciles rather than k grades

- The **Modified HL Test** (Phibbs, et. al. [1991])

  A Chi-Square test where deciles are defined in terms of the expected number of outcomes, rather than the number of observations in the grades

Measuring Accuracy and Precision
Mean-Squared Error

- Errors are made whenever decisions about an unknown quantity, such as PD, are based upon sample information.

- As we have seen, these errors will generally have two components:

  - some error may be due to **bias or inaccuracy**;

  - some error is due to **random variance or imprecision** arising from use of a sample

- A statistical measure that reflects both the accuracy and precision of an estimator is the Mean Squared Error of the Estimate (MSE):

MSE = Variance of the Estimate + Squared Bias of the Estimate

Example

Using MSE to Evaluate PD Rating System Granularity (Kiefer and Larson [2004])

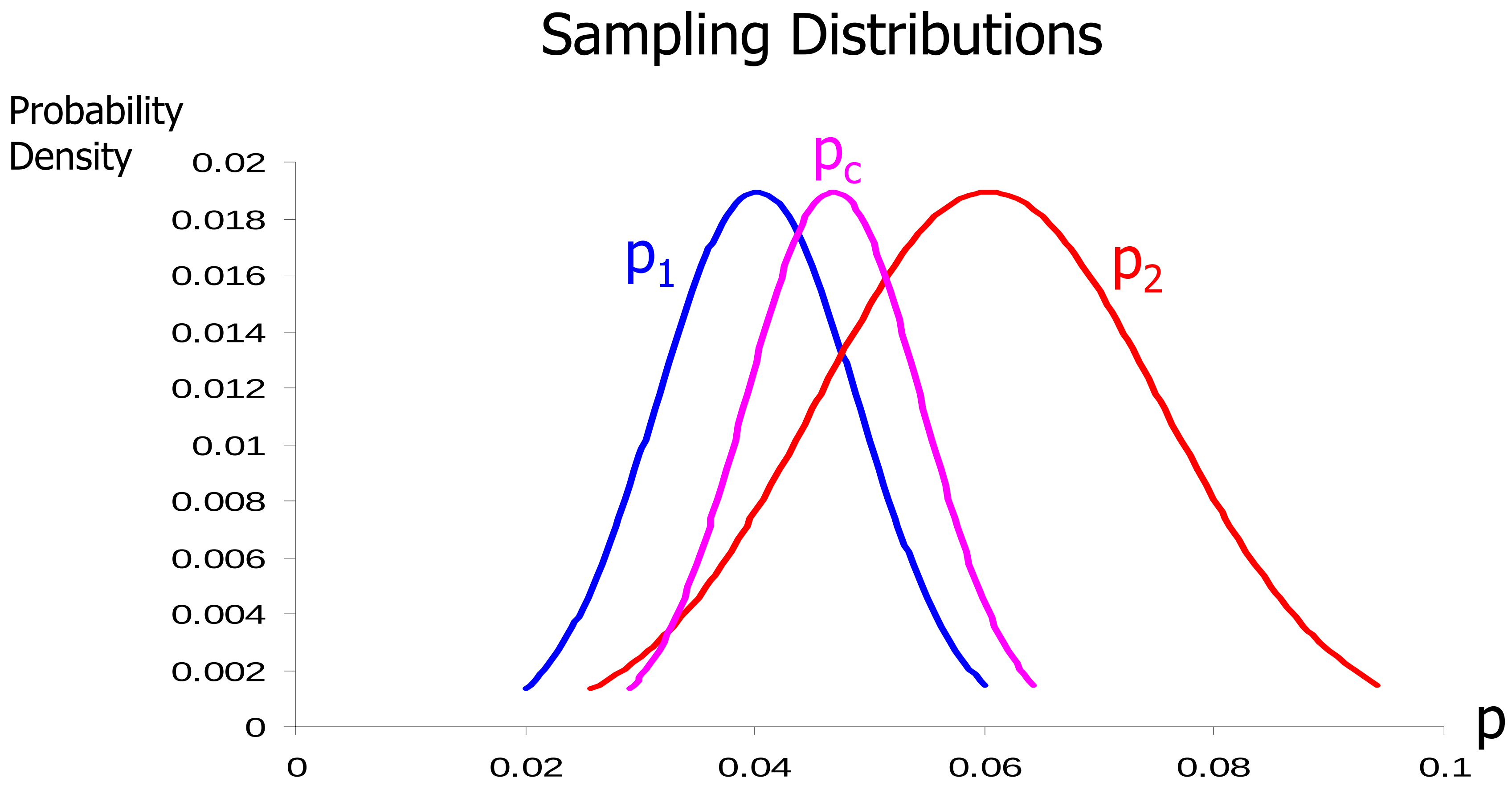
- Consider two different groups of obligors, with respective true (but unobservable) default rates given by θ_1=.04 and and θ_2=.06. We assume that defaults are uncorrelated.

- **We are interested in the question of whether these two groups should or should not be combined for the purposes of estimating default.**

- If we have n_1=500 and n_2=250 obligors from each group, we can we compute the sample default rates p_1 and p_2 to use as estimators of θ_1 and θ_2.

- Alternately, we could pool the sample data and estimate a single combined-group default rate, which we will call p_c.

Example: Using MSE



Example: Using MSE

Estimator	Expected Value	Bias		Variance	MSE $=$Variance$+$Bias^2
		θ_1 $=.04$	θ_2 $=.06$		
p1 $(n_1=500)$	Θ_1	0	n.a.	$Var(p_1)$ $=\theta_1(1-\theta_1)/n_1$ $=.0000768$	.0000768
p2 $(n_2=250)$	Θ_2	n.a.	0	$Var(p_2)$ $=\theta_2(1-\theta_2)/n_2$ $=.0002256$	.0002256
Portfolio with two rating buckets	Θ_1 and Θ_2	0	0	$Var(p_1)+Var(p_2)$ $=.0003024$	**.0003024**
p_c (Portfolio with one rating bucket)	$(n_1\theta_1+n_2\theta_2)/(n_1+n_2)$	$n_2(\theta_2-\theta_1)/(n_1+n_2)$ $=.0067$	$-n_1(\theta_2-\theta_1)/(n_1+n_2)$ $=-.0133$	$Var(p_c)$ $=(n_1\theta_1(1-\theta_1)+n_2\theta_2(1-\theta_2))/(n_1+n_2)^2$ $=.0000592$	**.0003406**

Since MSE from using the single combined estimate, p_c, is greater than the overall MSE from estimating p_1 and p_2 separately, the granularity is warranted from a perspective of minimizing errors in default rate estimation.

Conclusions

- Models can be built to different objectives
- Accuracy and precision are often required by the business use of a model
- Models should be evaluated in how they meet both design and use objectives
- Discriminatory power and forecast performance should both be assessed at the time of development and on a continuing basis subsequent to implementation.

References

- Basel Committee on Banking Supervision (BCBS), *International Convergence of Capital Measurement and Capital Standards: A Revised Framework,* updated November 2005.
- Hand, D.J. [2004] "Good Practice in Retail Scorecard Assessment", *OCC Credit Rating and Scoring Models Conference*, Washington, DC.
- Hosmer, D.W. and Lemeshow, S. [2000] *Applied Logistic Regression, Second Edition*. New York: John Wiley & Sons, Inc.
- Kiefer, N.M and Larson, C.E. [2004] "Evaluating Design Choices in Economic Capital Modeling: A Loss Function Approach" in *Economic Capital: A Practitioner Guide* (A. Dev, Editor). London: Risk Books.
- Pearson, K. [1900] "On the Criterion that a given System of Deviations from the Probable in the case of a Correlated System of Variables is such that it can be reasonably supposed to have arisen from Random Sampling." *Karl Pearson's Early Statistical Papers*. London: Cambridge University Press, 1956. pg. 339-357.
- Phibbs, C.S., Romano, P.S., Luft, H.S., Brown, B.W., Katz, P.P. [1991] "Improving the fit of logistic models for mortality and other rare events". *119th Annual Meeting of the American Public Health Association*, Atlanta, GA.
- Thomas, L.C., Edelman, D.B., and Crook, J.N. [2002] *Credit Scoring and it's Applications.* Philadelphia: SIAM