Goodness-of-Fit & ML Hypothesis Testing

Logit and Probit Models

CLABE 2025/2026

Marco Rosso

4 December 2025

Learning Goals

By the end of this lecture, you should be able to:

- **Derive** the maximum likelihood estimator (**MLE**) in a simple binomial model and interpret it as a sample proportion.
- **Explain why** the usual OLS R^2 is not appropriate for logit/probit models.
- **Define and interpret McFadden's pseudo-**R^2 and relate it to the log-likelihood.
- **Evaluate model performance** using classification tables, accuracy, sensitivity, and specificity.
- **Formulate and interpret LR, Wald, and LM** tests in ML frameworks.
- **Implement** these concepts in **Stata** (logit/probit, pseudo-R^2, classification, LR/Wald tests).

Example: Probability of Passing the Exam

Data Summary (by exam round):

Exam Round	Students	Passed	$\hat{\pi}$
June (1st)	95	71	≈ 0.747
July	38	21	≈ 0.553
September	27	10	≈ 0.370
February	15	12	≈ 0.800

Natural Questions:

- Does exam difficulty differ across rounds?
- Are better-prepared students taking the June exam?
- How confident are we in each estimate $\hat{\pi}$?

Modeling: Bernoulli Distribution

Setup:

- For each student i: $Y_i \in \{0, 1\}$ (fail or pass)
- All students in an exam round have same passing probability: $P(Y_i = 1) = \pi$
- Students are independent: $(Y_1, \ldots, Y_n)$ are i.i.d. Bernoulli(π)

Probability mass function for each observation:

$$P(Y_i = y_i | \pi) = \pi^{y_i}(1 - \pi)^{1-y_i}$$

where $y_i = 1$ if student passes, $y_i = 0$ if student fails.

Likelihood Function

Joint probability of observing the entire sample:

$$L(\pi; y_1, \ldots, y_n) = \prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{1-y_i}$$

Log-Likelihood (easier to work with):

$$\ell(\pi; y) = \log L(\pi; y) = \sum_{i=1}^{n} [y_i \log \pi + (1-y_i)\log(1-\pi)]$$

Why log-likelihood?

- Avoids numerical underflow (products of small probabilities)
- Sums are easier to optimize than products
- Derivatives (scores) are simpler

Maximum Likelihood Estimator (MLE)

Goal: Find $\hat{\pi}$ that maximizes $\ell(\pi; y)$.

First-order condition (Score = 0):

$$\frac{\partial \ell(\pi; y)}{\partial \pi} = \frac{\sum_{i=1}^{n} y_i}{\pi} - \frac{\sum_{i=1}^{n}(1 - y_i)}{1 - \pi} = 0$$

Let $n_1 = \sum_{i=1}^{n} y_i$ (number of students who passed). Then:

$$\frac{n_1}{\pi} = \frac{n - n_1}{1 - \pi}$$

Solving:

$$\boxed{\hat{\pi}_{\text{ML}} = \frac{n_1}{n} = \bar{y}}$$

Conclusion: The MLE of π is just the sample proportion!

Deriving the MLE: Analytical Solution (1)

Why showing the derivation?

- In this simple model, we can solve for $\hat{\pi}$ **analytically** (by hand).
- This illustrates the general principle: finding parameters where the gradient is zero.

Step-by-Step Solution:

0. **Given the log-likelihood:**

$$\ell(\pi; y) = \sum_{i=1}^{n} [y_i \log \pi + (1 - y_i) \log(1 - \pi)]$$

1. **Step 1: Take the first derivative (Score function):**

$$s(\pi) = \frac{\partial \ell(\pi; y)}{\partial \pi} = \sum_{i=1}^{n} \left[\frac{y_i}{\pi} - \frac{1 - y_i}{1 - \pi} \right]$$

Deriving the MLE: Analytical Solution (2)

2. **Step 2: Rearrange using $n_1 = \sum y_i$ and $n_0 = n - n_1$:**

$$s(\pi) = \frac{n_1}{\pi} - \frac{n_0}{1-\pi}$$

3. **Step 3: Set $s(\pi) = 0$ and solve for $\hat{\pi}$:**

$$\frac{n_1}{\pi} - \frac{n_0}{1-\pi} = 0 \quad \Rightarrow \quad \frac{n_1}{\pi} = \frac{n_0}{1-\pi} \quad \Rightarrow \quad n_1(1-\pi) = n_0\pi \quad \Rightarrow$$

$$\Rightarrow \quad n_1 - n_1\pi = n_0\pi \quad \Rightarrow \quad n_1 = (n_0 + n_1)\pi \quad \Rightarrow \quad \pi = \tfrac{n_1}{n_0+n_1} \quad \Rightarrow \quad \boxed{\hat{\pi} = \frac{n_1}{n}}$$

Crucial Difference with Logit/Probit: In Logit/Probit, the Score equations are non-linear and **cannot be solved by hand**. Software must use iterative approximation (Newton-Raphson) to find the maximum.

When is this Maximum? (Second-Order Condition)

Second derivative (Hessian):

$$H(\pi) = \frac{\partial^2 \ell(\pi; y)}{\partial \pi^2} = -\frac{n_1}{\pi^2} - \frac{n_0}{(1-\pi)^2} < 0 \quad \text{always}$$

Since $H(\pi) < 0$ everywhere:

- The log-likelihood is concave
- There is a unique maximum
- $\hat{\pi} = n_1/n$ is indeed the MLE

Intuition: The second derivative measures how "peaked" the log-likelihood is. Negative everywhere means we have a unique, stable maximum.

Computing the Hessian: Analytical Solution (1)

Goal: Verify that $H(\pi) < 0$ by computing the second derivative from scratch.

Step-by-Step Solution:

0. **Starting with the Score (first derivative):**

$$s(\pi) = \frac{\partial \ell(\pi; y)}{\partial \pi} = \frac{n_1}{\pi} - \frac{n_0}{1-\pi}$$

   where $n_1 = \sum y_i$ and $n_0 = n - n_1$.

1. **Step 1: Take the derivative of each term**

$$\begin{aligned} H(\pi) = \frac{\partial s(\pi)}{\partial \pi} &= \frac{\partial}{\partial \pi}\left(\frac{n_1}{\pi}\right) + \frac{\partial}{\partial \pi}\left(-\frac{n_0}{1-\pi}\right) \\ &= -\frac{n_1}{\pi^2} + \frac{\partial}{\partial \pi}\left(\frac{n_0}{1-\pi}\right) \end{aligned}$$

Computing the Hessian: Analytical Solution (2)

2. **Step 2: Apply the quotient rule to the second term**

   For $\frac{n_0}{1-\pi}$: the numerator is constant, so

$$\frac{\partial}{\partial \pi}\left(\frac{n_0}{1-\pi}\right) = n_0 \cdot \frac{\partial}{\partial \pi}(1-\pi)^{-1} = n_0 \cdot (-1)(1-\pi)^{-2} \cdot (-1) = \frac{n_0}{(1-\pi)^2}$$

3. **Step 3: Combine Step 1 and Step 2**

$$H(\pi) = -\frac{n_1}{\pi^2} - \frac{n_0}{(1-\pi)^2}$$

✓ **Result:** Since $n_1 > 0$, $n_0 > 0$, $\pi \in (0, 1)$:

$$\boxed{H(\pi) = -\frac{n_1}{\pi^2} - \frac{n_0}{(1-\pi)^2} < 0 \quad \text{always}}$$

Conclusion: The log-likelihood is strictly concave $\Rightarrow$ unique maximum at $\hat{\pi}$.

Example Calculation

June exam: $n = 95$ students, $n_1 = 71$ passed

$$\hat{\pi}_{\text{June}} = \frac{71}{95} \approx 0.747$$

September exam: $n = 27$ students, $n_1 = 10$ passed

$$\hat{\pi}_{\text{Sept}} = \frac{10}{27} \approx 0.370$$

Interpretation:

- June exam: students pass with probability $\approx 75\%$ (easier exam)
- September exam: students pass with probability $\approx 37\%$ (harder exam)
- Difference is substantial: might indicate different student ability or exam difficulty

Confidence Intervals from ML

Standard error of $\hat{\pi}$:

$$\text{SE}(\hat{\pi}) = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

95% Confidence Interval:

$$\hat{\pi} \pm 1.96 \cdot \text{SE}(\hat{\pi})$$

Example (June exam):

$$\text{SE} = \sqrt{\frac{0.747 \times 0.253}{95}} \approx 0.044$$

$$95\% \text{ CI} = [0.747 - 1.96(0.044),\ 0.747 + 1.96(0.044)] = [0.661, 0.833]$$

Example (September exam, smaller n):

$$\text{SE} = \sqrt{\frac{0.370 \times 0.630}{27}} \approx 0.093$$

$$95\% \text{ CI} = [0.370 - 1.96(0.093),\ 0.370 + 1.96(0.093)] = [0.188, 0.552]$$

Notice: **Wider CI with smaller n** (more uncertainty).

Connection to Logit/Probit

In the binomial model we just studied:

$$P(Y_i = 1) = \pi \quad \text{(constant for all students)}$$

In **logit/probit models**:

$$P(Y_i = 1 \mid X_i) = F(X_i\boldsymbol{\beta})$$

where:

- $F(\cdot)$ is a CDF (logistic or normal)
- The probability *varies* with observable characteristics X_i
- $\boldsymbol{\beta}$ is the parameter vector (like π before, but multidimensional)

Key similarities:

- Log-likelihood: $\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i \log F(X_i\boldsymbol{\beta}) + (1 - y_i) \log(1 - F(X_i\boldsymbol{\beta}))]$
- MLE: maximizes $\ell(\boldsymbol{\beta})$
- Inference: uses scores, Hessian, LR/Wald/LM tests

From Binomial Model to Model Comparison (1)

What we learned so far:

- We can estimate π (a single parameter) via maximum likelihood
- At the MLE: $\hat{\pi} = \bar{y}$ (sample mean)
- The log-likelihood tells us how good the fit is

Now in logit/probit:

- Instead of one π, we have many parameters $\boldsymbol{\beta}$
- The MLE $\hat{\boldsymbol{\beta}}$ maximizes $\ell(\boldsymbol{\beta})$
- Different models have *different log-likelihoods*

From Binomial Model to Model Comparison (2)

Key insight for model comparison:

- **Constant-only model:** Only intercept (no regressors) $\Rightarrow \ell_{\text{const}}$
- **Full model:** All regressors included $\Rightarrow \ell_{\text{full}}$
- **Always:** $\ell_{\text{full}} \geq \ell_{\text{const}}$ (adding variables can't make fit worse)
- The **difference** in log-likelihoods measures improvement

The Problem: R^2 in Nonlinear Models

OLS: $R^2 = 1 - \frac{\sum \hat{u}_i^2}{\sum (y_i - \bar{y})^2}$ works great

Logit/Probit: R^2 is **inappropriate** because:

1. Dependent variable is binary $\Rightarrow$ fixed variation
2. Predictions $\hat{P}_i \in (0, 1)$ are probabilities, not 0 or 1
3. Residuals $\hat{u}_i = y_i - \hat{P}_i$ are not identically distributed

Solution: Use alternatives:

- Pseudo-R^2 (based on likelihoods)
- Classification accuracy (based on predictions)
- ML hypothesis tests (LR, Wald, LM)

Log-Likelihood for Binary Models

For observation i:

$$P(Y_i = y_i|\mathbf{X}_i) = F(\mathbf{X}_i\boldsymbol{\beta})^{y_i} \cdot [1 - F(\mathbf{X}_i\boldsymbol{\beta})]^{1-y_i}$$

Sum over all n observations:

$$\ell(\boldsymbol{\beta}; \mathbf{y}) = \sum_{i=1}^{n} [y_i \ln F(\mathbf{X}_i\boldsymbol{\beta}) + (1 - y_i) \ln(1 - F(\mathbf{X}_i\boldsymbol{\beta}))] \tag{1}$$

Key facts:

- $\ell(\boldsymbol{\beta}; \mathbf{y}) < 0$ always (log of probabilities)
- **Larger (less negative)** $=$ better fit
- **Logit:** $F(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$
- **Probit:** $F(z) = \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt$ (std normal)

McFadden's Pseudo-R^2

The most common measure:

$$R^2_{\text{McF}} = 1 - \frac{\ell_{\text{full}}}{\ell_{\text{const}}} \tag{2}$$

where:

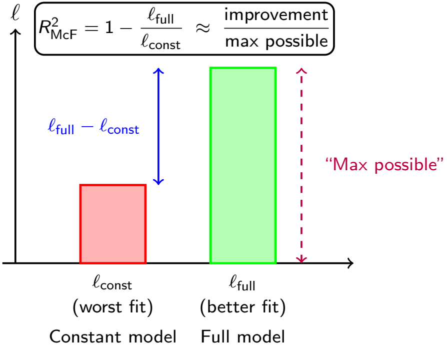
- ℓ_{full} = log-likelihood with all regressors
- ℓ_{const} = log-likelihood with only constant

Interpretation:

- Ranges from 0 to 1 (approximately)
- $R^2_{\text{McF}} = 0.15$: 15% improvement over constant-only model
- **Benchmark:** 0.2–0.4 is "good" for discrete choice
- **NOT comparable to OLS** R^2 (tends to be much lower)

From Log-Likelihood Difference to McFadden's R^2

Core idea: Compare how much the log-likelihood improves when we move from the constant-only model to the full model.



Interpretation:

- If the full model does not improve the log-likelihood $\Rightarrow R^2_{\text{McF}} \approx 0$.
- Larger improvements in log-likelihood $\Rightarrow$ higher R^2_{McF}.

Why McFadden's R^2 Makes Sense

Comparison with OLS R^2:

Metric	OLS	Logit/Probit
Numerator	$\sum \hat{u}_i^2$	ℓ_{full}
Denominator	$\sum (y_i - \bar{y})^2$	ℓ_{const}
Intuition	% of variance explained	% of likelihood improved
Formula	$R^2 = 1 - \frac{\text{SSR}}{\text{SST}}$	$R^2_{\text{McF}} = 1 - \frac{\ell_{\text{full}}}{\ell_{\text{const}}}$

Why not use OLS R^2 for logit/probit?

1. Binary $y \in \{0, 1\}$ has **fixed variation** (no natural "total")
2. Predictions $\hat{P}_i \in (0, 1)$ are probabilities (not errors)
3. Residuals $\hat{u}_i = y_i - \hat{P}_i$ are **heteroskedastic by design**
4. Log-likelihood is the natural scale for likelihood-based models

McFadden's R^2: Practical Guidelines

Benchmark interpretation (from literature):

Range	Interpretation
$0.0 - 0.1$	Weak fit (consider different model)
$0.1 - 0.2$	Fair fit (acceptable for exploratory work)
$0.2 - 0.4$	**Good fit (typical for discrete choice)**
$0.4 - 0.6$	Very good fit (model does well)
> 0.6	Excellent fit (rare in practice)

Important note:

- McFadden R^2 is **NOT comparable to OLS** R^2
- OLS R^2 tends to be much *higher* than McFadden
- Example: OLS $R^2 = 0.7 \approx$ McFadden $R^2 = 0.15$ (same model!)
- Always report **McFadden + classification accuracy** together

Alternative Pseudo-R^2 Measures

- **Cox-Snell:** $R^2_{\text{CS}} = 1 - \left(\frac{L_{\text{const}}}{L_{\text{full}}}\right)^{2/n}$ (uses likelihood, not log-likelihood)
- **Nagelkerke:** $R^2_{\text{Nag}} = \frac{R^2_{\text{CS}}}{R^2_{\text{CS,max}}}$ (scales to [0,1])
- **Information Criteria:**
  - $\text{AIC} = 2k - 2\ell(\hat{\beta})$ (Smaller is better)
  - $\text{BIC} = k\ln(n) - 2\ell(\hat{\beta})$ (More penalty on k)

In practice: Report McFadden's R^2 + classification accuracy

Classification Table (Confusion Matrix)

Process:

1. Compute predicted probability: $\hat{P}_i = F(\mathbf{X}_i\hat{\beta})$
2. Apply threshold (typically 0.5):

$$\hat{y}_i = \begin{cases} 1 & \text{if } \hat{P}_i > 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

3. Tabulate against actual outcomes

		Predicted	
		$\hat{y} = 0$	$\hat{y} = 1$
Actual	$y = 0$	TN	FP
	$y = 1$	FN	TP

Accuracy Metrics

Accuracy (Overall):

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{n}$$

$\rightarrow$ Proportion of correct predictions

Sensitivity (True Positive Rate):

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$\rightarrow$ Proportion of actual 1's correctly predicted

Specificity (True Negative Rate):

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$\rightarrow$ Proportion of actual 0's correctly predicted

Trade-off: Lower threshold $\Rightarrow$ higher sensitivity, lower specificity

Example: Ketchup Brand Choice (Model A)

Logit with pricebrand1 only

	Predicted		
brand1	0	1	Total
0	2488	0	2488
1	310	0	310
Total	2798	0	2798

Analysis:

- Accuracy: $2488/2798 = 88.9\%$
- Sensitivity: $0/310 = 0\%$ (predicts everyone as 0!)
- Specificity: $2488/2488 = 100\%$

Problem: Model predicts everyone chooses $y = 0$. Weak predictive power.

Example: Ketchup Brand Choice (Model B)

Logit with pricebrand1 AND pricebrand2

	Predicted		
brand1	0	1	Total
0	2465	23	2488
1	309	1	310
Total	2774	24	2798

Improved metrics:

- Accuracy: $(2465 + 1)/2798 = 88.1\%$ (worse overall)
- Sensitivity: $1/310 = 0.32\%$ (still very low)
- Specificity: $2465/(2465 + 23) = 99.1\%$ (better)

Insight: Imbalanced data (mostly $y = 0$) limits predictive ability

The “Trinity” of ML Tests

Testing $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{q}$ vs $H_1 : \mathbf{R}\boldsymbol{\beta} \neq \mathbf{q}$

Test	**Formula**	**Needs**	**When**
LR	$-2(\ell_R - \ell_U)$	Both	Comparing models
Wald	(see next)	Unrestr. only	Easy to estimate
LM	(see next)	Restr. only	Rare in practice

Distribution: All $\sim \chi^2_q$ under H_0, where $q = \#$ restrictions

Decision: If Test Stat $> \chi^2_{q,\alpha} \Rightarrow$ **Reject** H_0

Likelihood Ratio (LR) Test

Formula:

$$\text{LR} = -2\left(\ell_R - \ell_U\right) \tag{4}$$

where ℓ_R = restricted model, ℓ_U = unrestricted model

Intuition:

- If H_0 true: $\ell_R \approx \ell_U \Rightarrow \text{LR} \approx 0 \Rightarrow$ don't reject
- If H_0 false: $\ell_U \gg \ell_R \Rightarrow$ LR large $\Rightarrow$ reject

Advantages:

- Simple to compute (just log-likelihoods)
- Easy to interpret
- Standard for comparing nested models

LR Test Example

Setup: Test $H_0 : \beta_{\text{pricebrand2}} = 0$

- Restricted: logit with pricebrand1 only $\Rightarrow \ell_R = -836.30363$
- Unrestricted: logit with pricebrand1 + pricebrand2 $\Rightarrow \ell_U = -715.62915$

Calculation:

$$\text{LR} = -2(-836.30363 - (-715.62915)) = -2 \times (-120.67448) = 241.35 \tag{5}$$

Critical value ($q = 1$, $\alpha = 0.05$): $\chi^2_{1,0.05} = 3.841$

Decision: $241.35 > 3.841 \Rightarrow$ **Reject** H_0

p-value: $\Pr(\chi^2_1 > 241.35) \approx 0.000$

Wald Test

Formula:

$$\text{Wald} = (\mathbf{R}\hat{\beta}_U - \mathbf{q})^\top [\mathbf{R}\widehat{\text{Var}}(\hat{\beta}_U)\mathbf{R}^\top]^{-1}(\mathbf{R}\hat{\beta}_U - \mathbf{q}) \tag{6}$$

$\rightarrow$ Under H_0: Wald $\sim \chi^2_q$

In Stata just use `test` command

```
logit brand1 pricebrand1 pricebrand2
test pricebrand2 = 0
```

Advantages:

- Only need unrestricted model
- Easiest to implement in practice
- Asymptotically equivalent to LR

Lagrange Multiplier (LM) Test

Formula:

$$\text{LM} = \mathbf{s}(\hat{\beta}_R)^\top [\mathbf{H}(\hat{\beta}_R)]^{-1} \mathbf{s}(\hat{\beta}_R) \tag{7}$$

where $\mathbf{s}$ = score (gradient), $\mathbf{H}$ = Hessian

When to use:

- Restricted model is easy to estimate
- Unrestricted model is complex
- *Rarely used in practice* for binary choice

Asymptotic equivalence: LR, Wald, and LM all $\sim \chi^2_q$ as $n \to \infty$

Practical Workflow: Step 1 - Load Data & Estimate Model

```
clear all
use "ketchup_binary_clabe.dta"

* Baseline model
logit brand1 pricebrand1
est store model_A
scalar ll_A = e(ll)
```

Output: $\ell_A = -836.30363$

Practical Workflow: Step 2 - Compute McFadden's R^2

```
* Constant-only model
logit brand1
scalar ll_const = e(ll)

* McFadden's R-squared
scalar mcfadden = 1 - (ll_A / ll_const)
display "McFadden's R-sq = " mcfadden
```

Example:

- $\ell_{\text{const}} = -1086.1163$
- $\ell_A = -836.30363$
- $R^2_{\text{McF}} = 1 - (-836.30363 / -1086.1163) = 0.230$

Practical Workflow: Step 3 - Compute Predicted Probabilities

```
* Re-estimate model A
logit brand1 pricebrand1

* Predicted probability (logit formula)
gen phat_A = exp(_b[_cons] + _b[pricebrand1]*pricebrand1) / ///
             (1 + exp(_b[_cons] + _b[pricebrand1]*pricebrand1))

* Classification (0.5 threshold)
gen yhat_A = (phat_A > 0.5)

* Classification table
tab brand1 yhat_A
```

Practical Workflow: Step 4 - Extended Model & LR Test

```
* Unrestricted model
logit brand1 pricebrand1 pricebrand2
est store model_B
scalar ll_B = e(ll)

* LR test (manual)
scalar LR = -2 * (ll_A - ll_B)
scalar df = 1
scalar crit = invchi2(df, 0.95)
scalar pval = chi2tail(df, LR)

display "LR = " LR
display "Critical value = " crit
display "P-value = " pval
```

Practical Workflow: Step 5 - Wald Test

```
* Already estimated model B
logit brand1 pricebrand1 pricebrand2

* Single restriction
test pricebrand2 = 0

* Joint restrictions
test pricebrand1 pricebrand2

* Built-in LR test
lrtest model_A model_B
```

Key Takeaways

Goodness-of-Fit:

- ✓ McFadden's $R^2 = 1 - \frac{\ell_{\text{full}}}{\ell_{\text{const}}}$ (benchmark: 0.2–0.4 is good)
- ✓ Classification accuracy (watch for imbalanced data)
- ✓ AIC/BIC for model comparison

Hypothesis Testing:

- ✓ LR, Wald, LM all asymptotically χ^2_q
- ✓ LR: simple (just log-likelihoods)
- ✓ Wald: easiest in practice (Stata `test` command)
- ✓ All three asymptotically equivalent

Practical Rule:

- ✓ Always report pseudo-R^2 + classification accuracy
- ✓ Manual computation = understanding
- ✓ Check imbalance in outcomes

Exam Checklist

- ☐ McFadden's R^2 formula
- ☐ Classification table (TN, TP, FN, FP)
- ☐ Accuracy, Sensitivity, Specificity
- ☐ LR test formula: $LR = -2(\ell_R - \ell_U)$
- ☐ Distribution: χ^2_q
- ☐ Decision rule: compare to critical value
- ☐ Logit CDF: $F(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$
- ☐ Probit CDF: $F(z) = \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt$
- ☐ Manual computation in Stata using scalars
- ☐ Difference: LR vs Wald vs LM

Questions?

Let's apply this in Stata