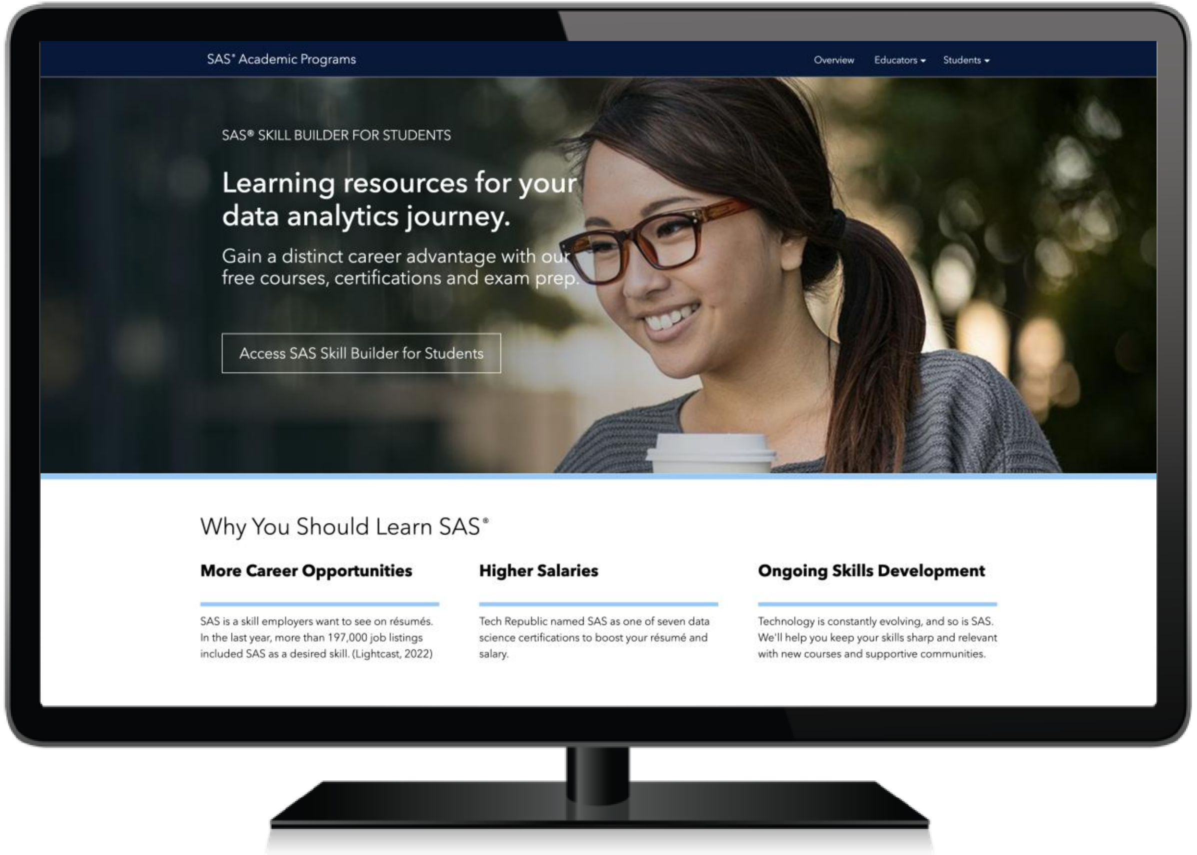
Categorical Data Analysis

Research Triangle SAS User Group

October 17, 2024

Tom Grant – Academic Training Consultant - tom.grant@sas.com

SAS Skill Builder for Students



SAS® SKILL BUILDER FOR STUDENTS

Learning resources for your data analytics journey.

Gain a distinct career advantage with our free courses, certifications and exam prep.

Why You Should Learn SAS®

More Career Opportunities

SAS is a skill employers want to see on résumés. In the last year, more than 197,000 job listings included SAS as a desired skill. (Lightcast, 2022)

Higher Salaries

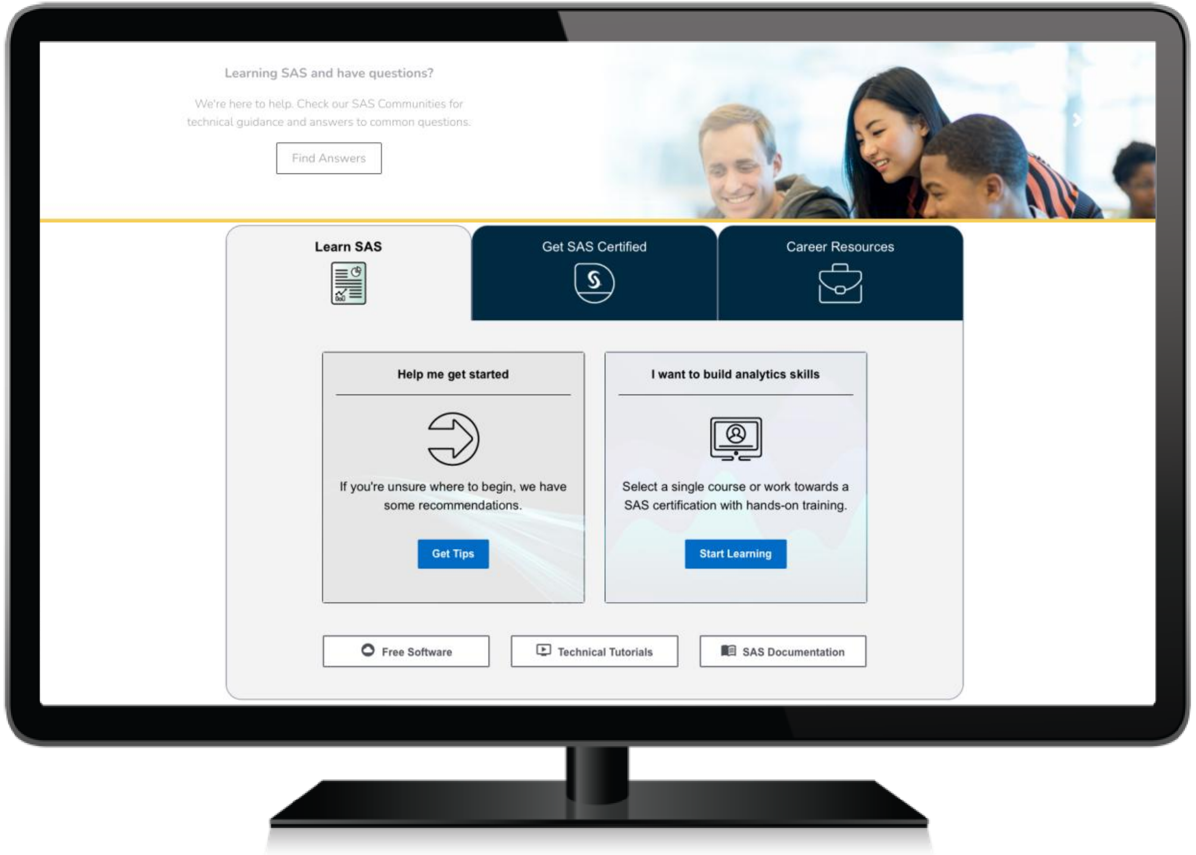
Tech Republic named SAS as one of seven data science certifications to boost your résumé and salary.

Ongoing Skills Development

Technology is constantly evolving, and so is SAS. We'll help you keep your skills sharp and relevant with new courses and supportive communities.

Web Experience - Example

SAS Educator Portal



Learning SAS and have questions?

We're here to help. Check our SAS Communities for technical guidance and answers to common questions.

Help me get started

If you're unsure where to begin, we have some recommendations.

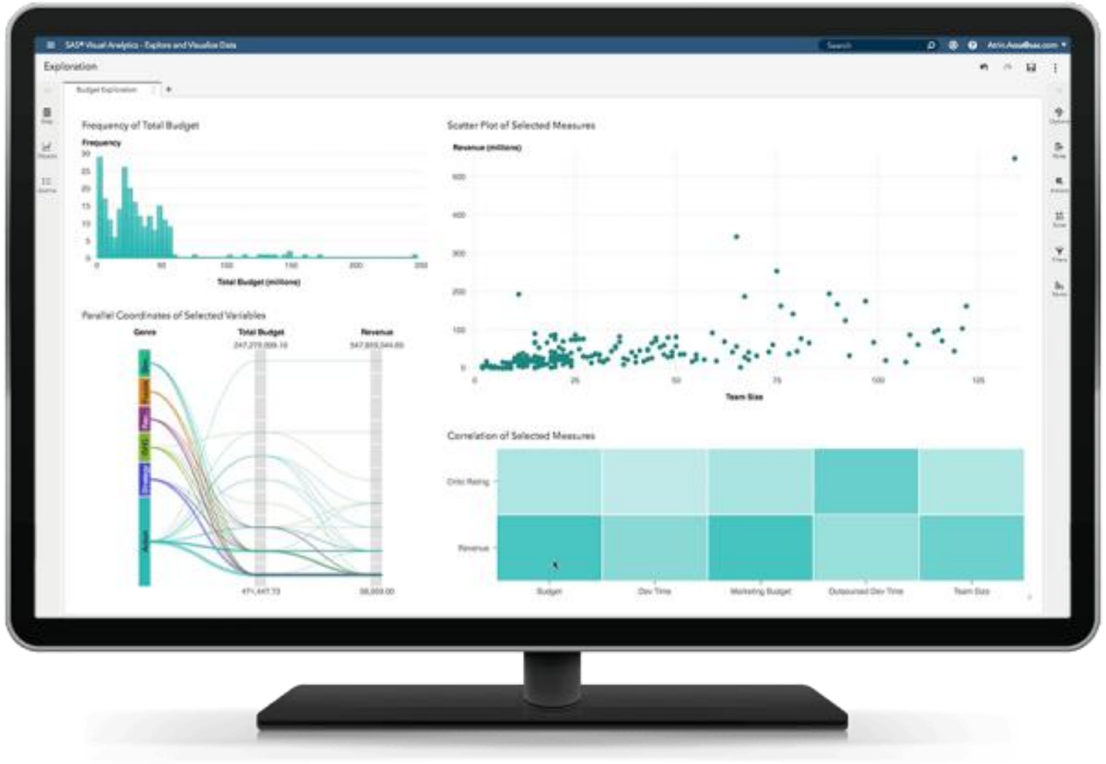
I want to build analytics skills

Select a single course or work towards a SAS certification with hands-on training.

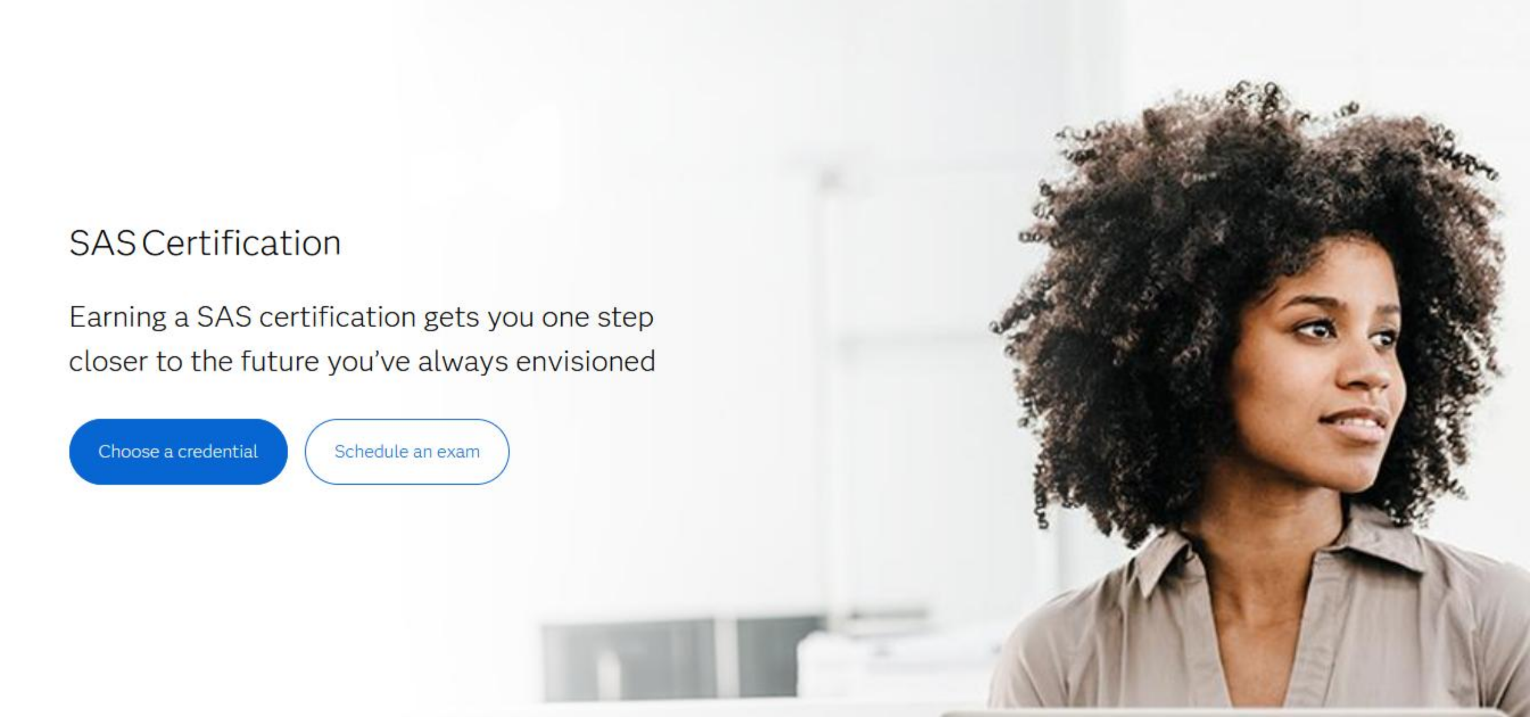
Virtual Learning Environment - Example

Free Academic Software: Viya Advanced

Getting Started with SAS Viya for Learners



SAS Certification



SAS Certification

Earning a SAS certification gets you one step closer to the future you've always envisioned

Categorical Data

- *Categorical data* represent categories, classes and classifications, groups, or qualitative characteristics or attributes.
  - respondent gender (**male** or **female**)
  - product disposition (**conforming** or **nonconforming**)
  - patient mortality (**survived** or **died**)
- *Continuous* data represent measurements.
  - length, time, temperature, concentration
- Categorical data are *qualitative*, continuous data are *quantitative*.
- Categorical data values are *discrete* and the distance between categories is unknown.

Frequency Table Analysis

- Frequency tables are useful because they can do the following:
  - help detect erroneous data points
  - can be used to assess associations among categorical variables
  - are helpful in determining where possible problems might occur in a logistic regression model

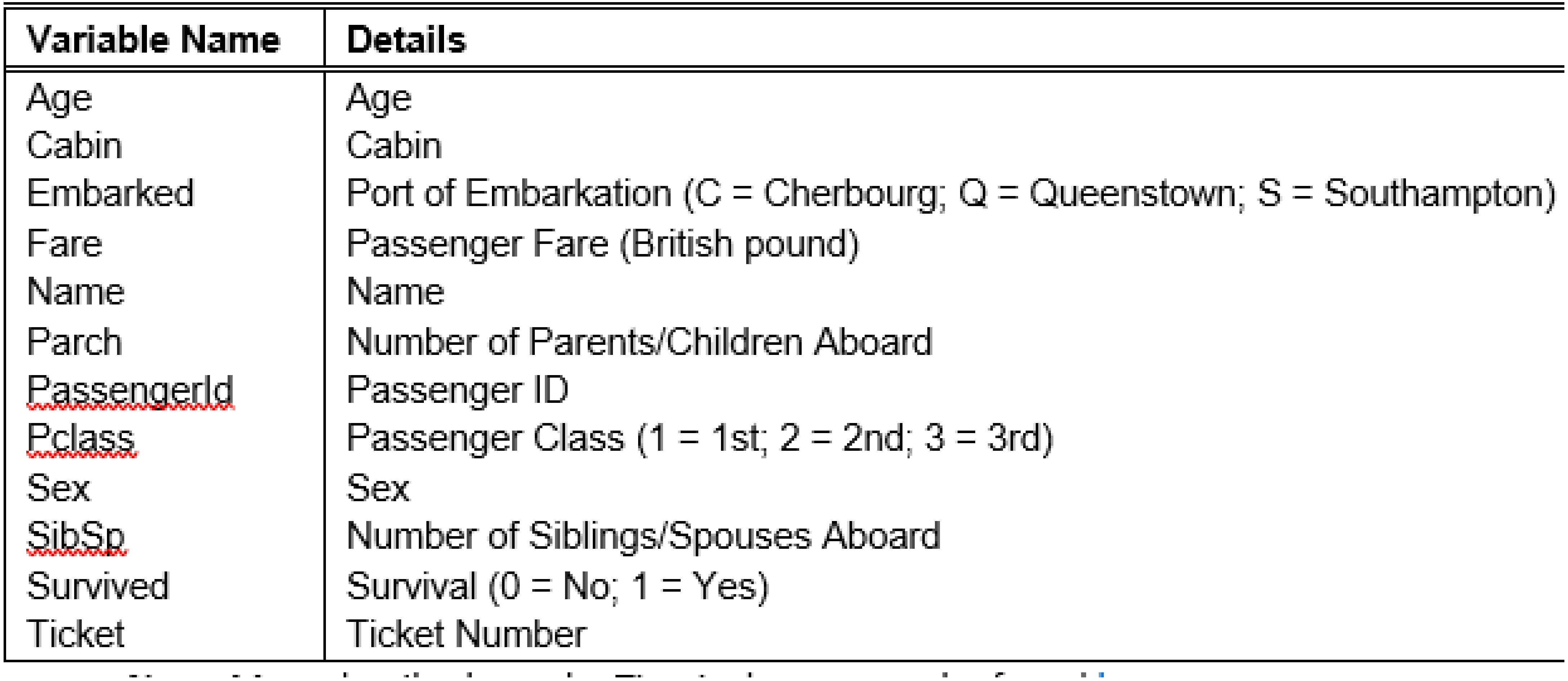
The FREQ Procedure

- General form of the FREQ procedure:

```
PROC FREQ DATA=SAS-data-set;
        TABLES table-requests </ options>;
RUN;
```

Titanic Insurance Co., Inc.

Data on Passengers



Variable Name	Details
Age	Age
Cabin	Cabin
Embarked	Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)
Fare	Passenger Fare (British pound)
Name	Name
Parch	Number of Parents/Children Aboard
PassengerId	Passenger ID
Pclass	Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
Sex	Sex
SibSp	Number of Siblings/Spouses Aboard
Survived	Survival (0 = No; 1 = Yes)
Ticket	Ticket Number

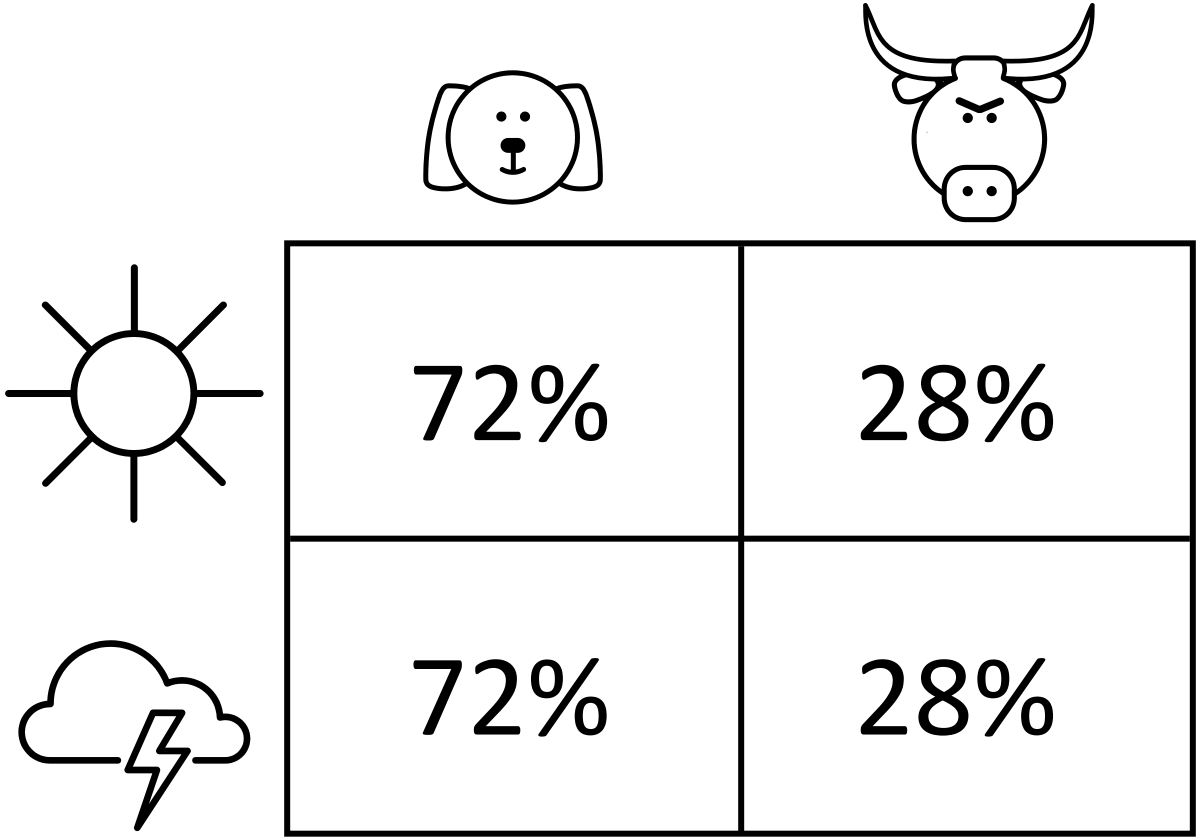
Demo – Titanic Data – Explore & Visualize

https://www.sas.com/en_us/learn/academic-programs.html

Categorical Variables Association

- An association exists between two categorical variables if the distribution of one variable changes when the level (or value) of the other variable changes.
- If there is no association, the distribution of the first variable is the same regardless of the level of the other variable.

No Association



	🐶	🐂
☀	72%	28%
⛈	72%	28%

Is your manager’s mood associated with the weather?

Association

	🐶	🐂
☀	82%	18%
⛈	60%	40%

Is your manager's mood associated with the weather?

Null Hypothesis

- There is ***no*** association between the weather and your boss's mood.
- The probability of your boss being in a good mood is the same on cloudy and sunny days.

- Alternative Hypothesis

- There ***is*** an association between weather and your boss's mood.
- The probability of your boss being in a good mood is ***not*** the same on cloudy and sunny days.

Chi-Square Test

NO ASSOCIATION

observed frequencies = expected frequencies

ASSOCIATION

observed frequencies ≠ expected frequencies

Note: The expected frequencies are calculated by this formula: (row total * column total) / sample size.

Chi-Square Tests

- Chi-square tests and the corresponding *p*-values can do the following:
  - determine whether an association exists
  - do not measure the strength of an association
  - depend on and reflect the sample size

$$\boxed{\chi^2 = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(Obs_{ij} - Exp_{ij})^2}{Exp_{ij}}}$$

Demo – Titanic Data – SAS Studio Task – Table Analysis

https://www.sas.com/en_us/learn/academic-programs.html

Odds Ratios

- An *odds ratio* indicates how much more likely, with respect to odds, a certain event occurs in one group relative to its occurrence in another group.

- Example: How do the odds of males surviving compare to those of females?

$$\boxed{\text{Odds} = \frac{p_{event}}{1 - p_{event}}}$$

Probability versus Odds of an Outcome

	Outcome		
	No	**Yes**	**Total**
Group A	20	60	80
Group B	10	90	100
Total	30	150	180

Probability of **Yes** in Group B = 0.90 ÷ Probability of **No** in Group B = 0.10

Odds of **Yes** in Group B = **0.90** ÷ **0.10** = **9**

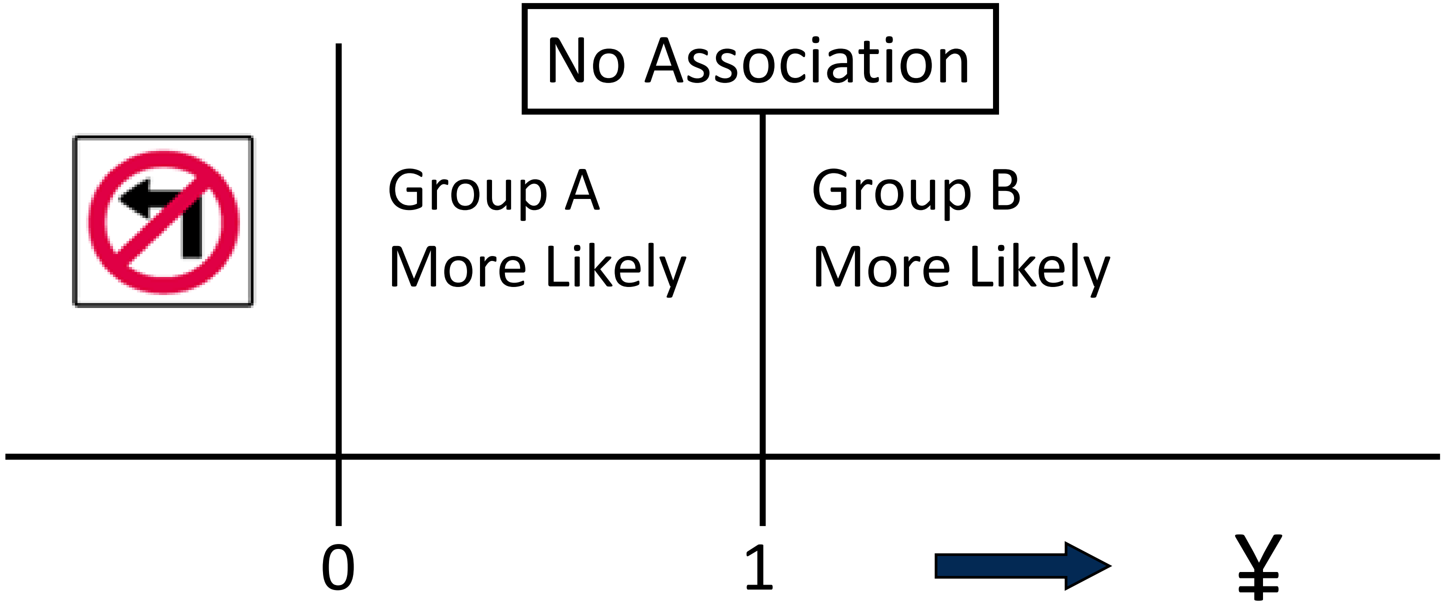
Odds Ratio

	Outcome		
	No	Yes	Total
Group A	20	60	80
Group B	10	90	100
Total	30	150	180

Odds of **Yes** in **Group B** = **9** ÷ Odds of **Yes** in **Group A** = **3**

Odds Ratio, **B** to **A** = **9** ÷ **3** = **3**

Properties of the Odds Ratio, B to A



No Association

Group A
More Likely

Group B
More Likely

0

1

¥

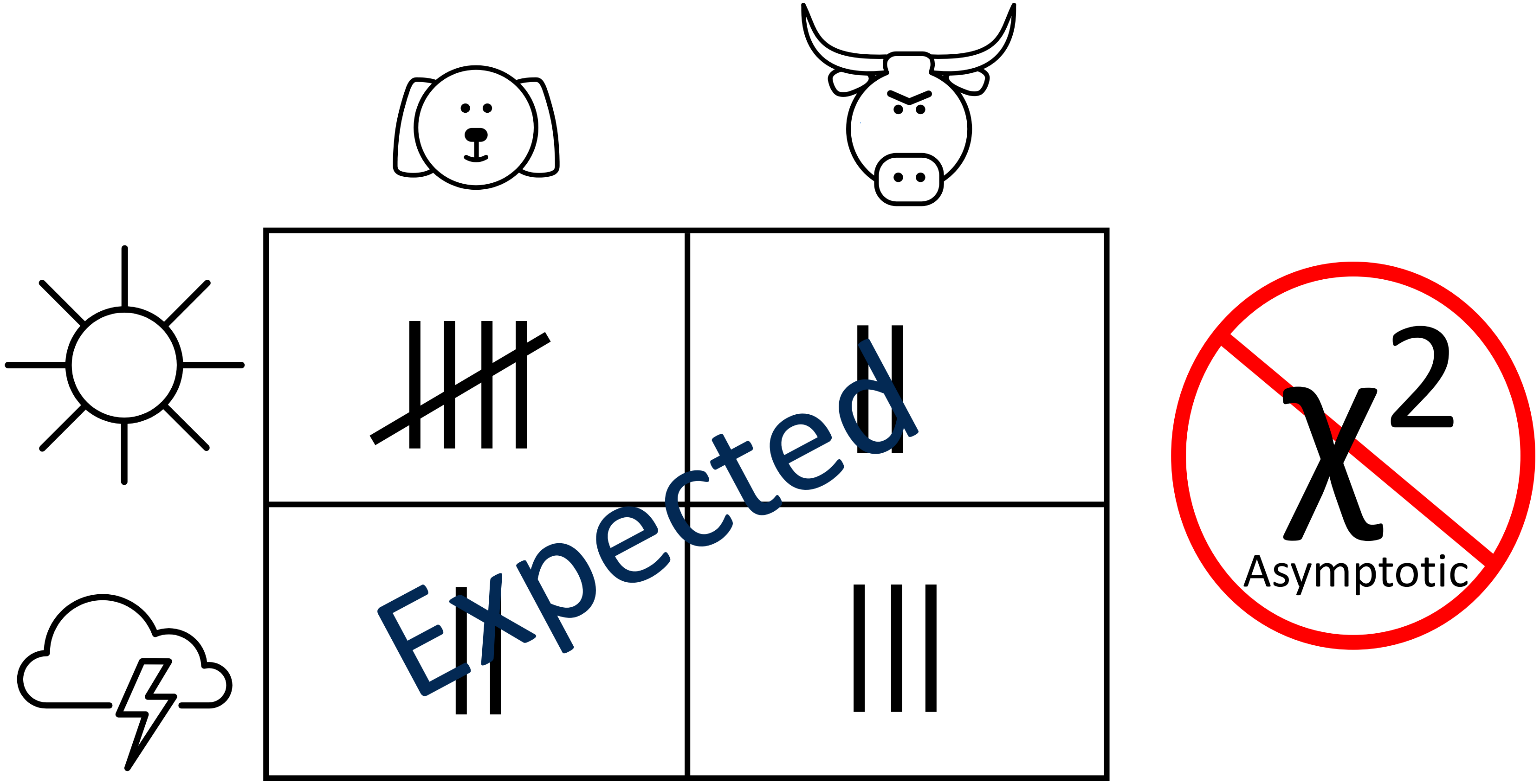
Tests of Association

Row Variable	Column Variable	R x C table	2 x 2 table
Ordinal	Ordinal	Mantel-Haenzel χ^2	CI for odds ratio
Nominal	Ordinal	Mean score Statistic	CI for odds ratio
Nominal	Nominal	Pearson χ^2	CI for odds ratio

Measures of Association Strength

Row Variable	Column Variable	R x C table	2 x 2 table
Ordinal	Ordinal	Spearman Correlation	Odds Ratio
Nominal	Ordinal	Uncertainty Coefficient c|r	Odds Ratio
Nominal	Nominal	Uncertainty Coefficient c|r	Odds Ratio

When Not to Use the Asymptotic χ^2



When more than 20% of cells have expected counts less than five

Demo – adding Odds Ratios

https://www.sas.com/en_us/learn/academic-programs.html

Logistic Regression

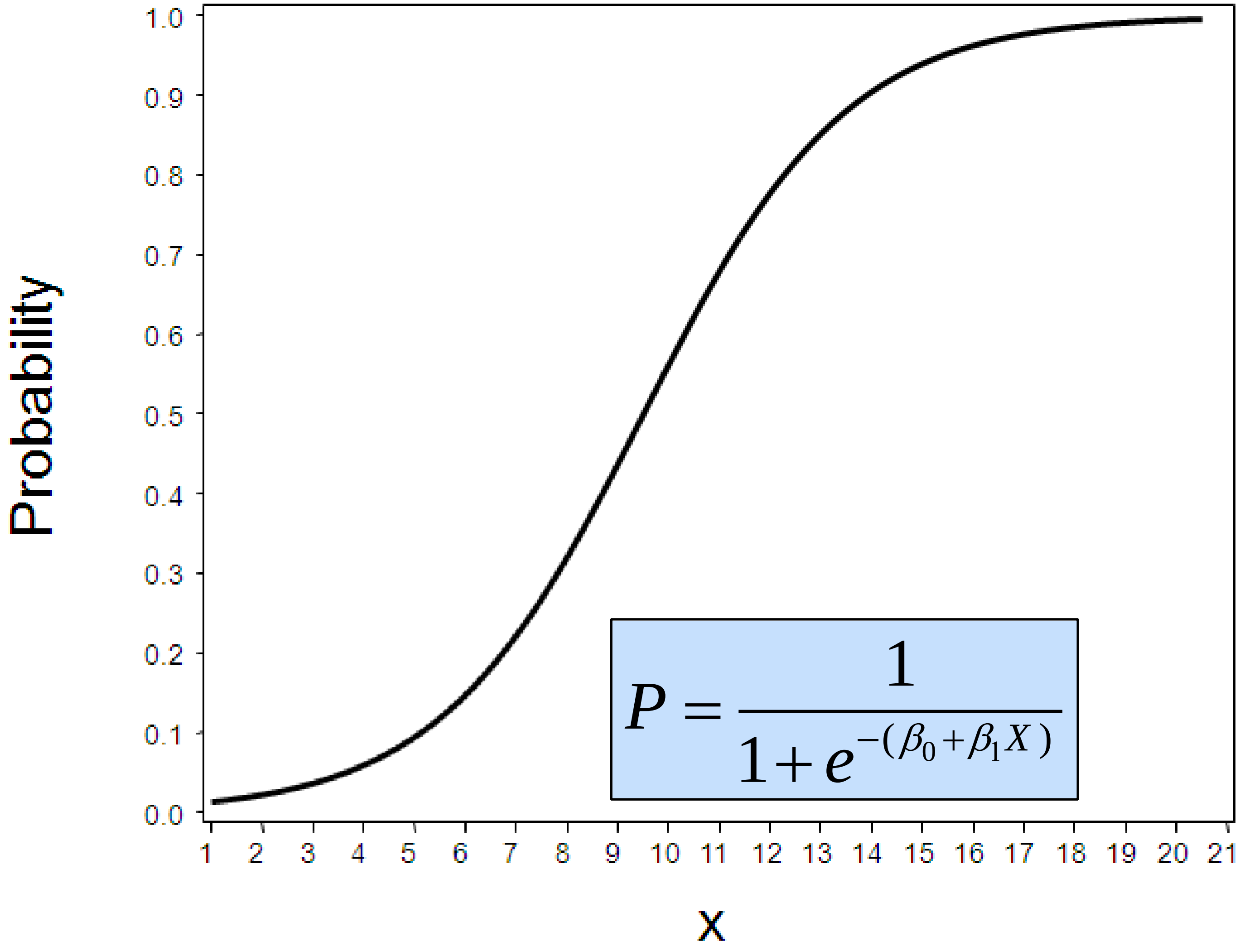
https://www.sas.com/en_us/learn/academic-programs.html

Why Not Ordinary Least Squares Regression?

$$\boxed{\textit{OLS Regression: } Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i}$$

-
  - The random error term ε has a normal distribution with a mean of zero.
  - The random error term has a constant variance.
  - The errors εi are independent.
  - The model is correctly specified.
- In logistic regression, the first two assumptions are violated. Therefore, OLS is not the best method for parameter estimation.

Logistic Regression Model



$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

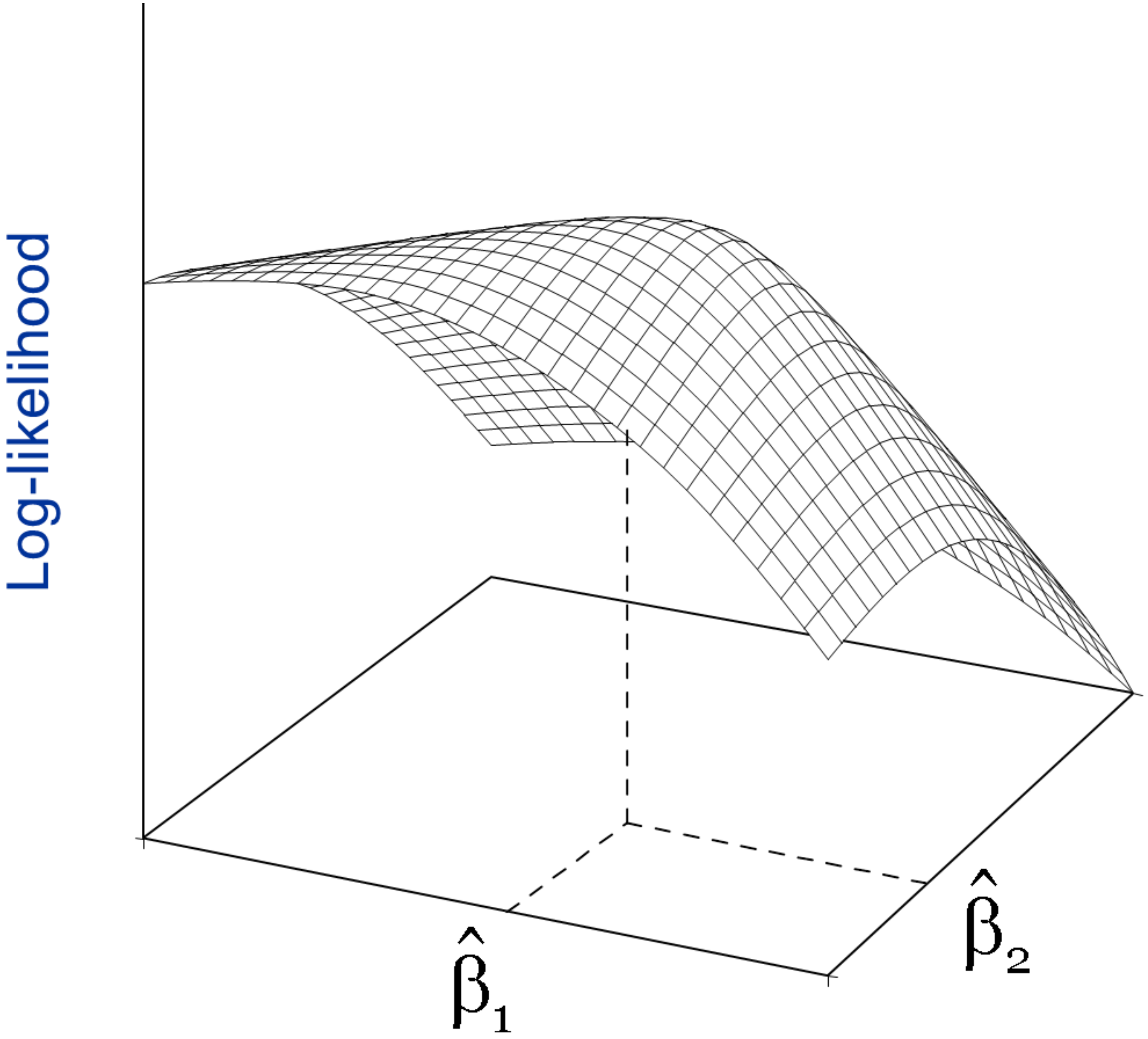
Logit Transformation

- Logistic regression models transformed probabilities, called *logits**,
-

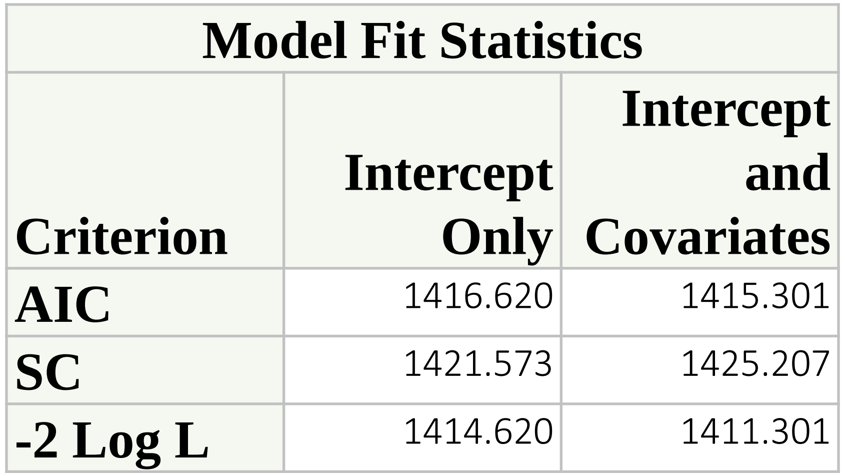
$$\text{logit}(p_i) = \ln\left(\frac{p_i}{(1-p_i)}\right) = \beta_0 + \beta_1 X$$

- where
- *i* indexes all cases (observations)
- p_i is the probability that the event (a sale, for example) occurs in the *i*th case
- ln is the natural log (to the base e).
- * The logit is the natural log of the odds.

Maximum Likelihood Estimation



Model Fit Statistics



Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	1416.620	1415.301
SC	1421.573	1425.207
-2 Log L	1414.620	1411.301

- Akaike’s information criterion (AIC)

$$AIC = -2Log\left(L\right) + 2k$$

- Schwarz Bayesian information criterion (SC)

$$SC = -2Log\left(L\right) + k\log\left(n\right)$$

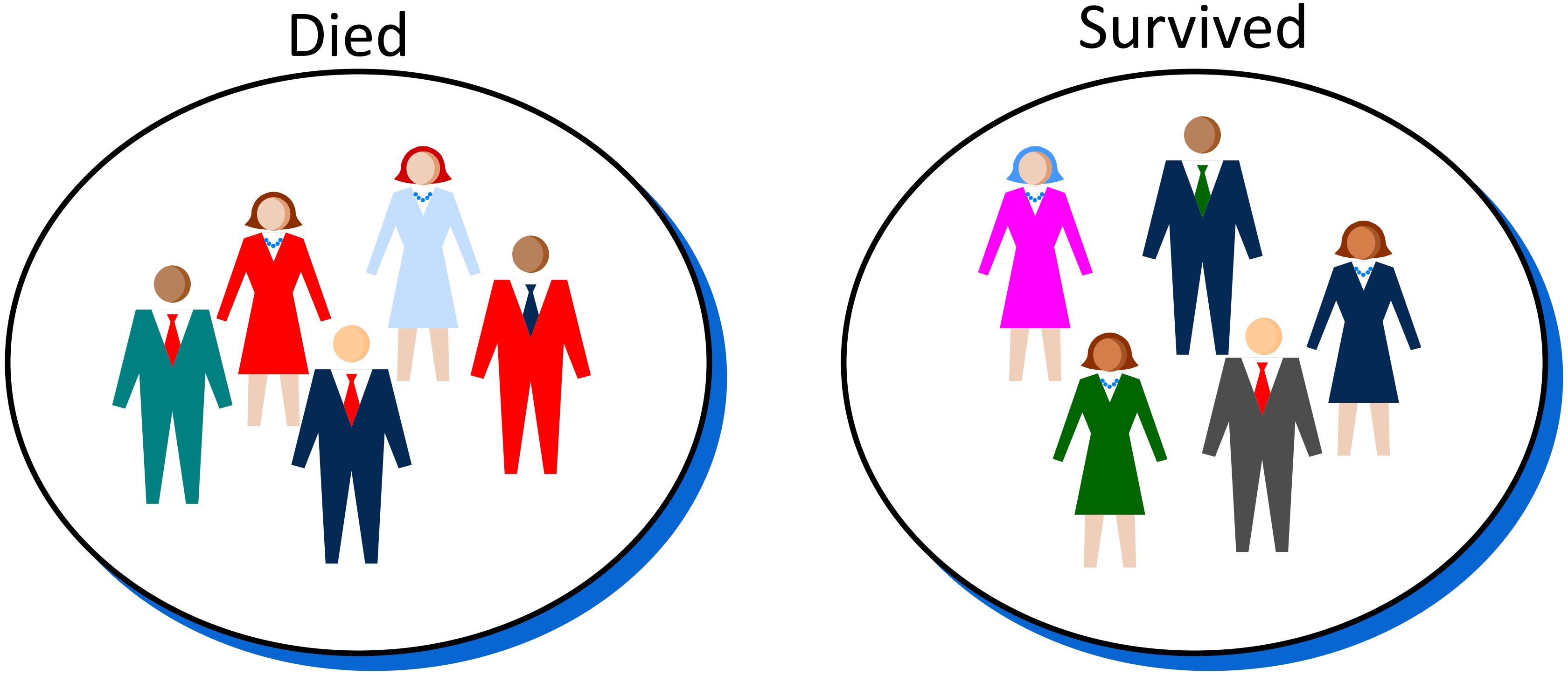
- Smaller values indicate a better model.

Predictive Accuracy

- Examining the percentage of concordant, discordant, and tied pairs is a way to assess the predictive accuracy of the model.
- In general, you want a high percentage of concordant pairs and a low percentage of discordant pairs.

Comparing Pairs

- To find concordant, discordant, and tied pairs, compare everyone who had the outcome of interest against everyone who did not.



Died

Survived

Concordant Pair

- Compare a 20-year-old who survived with a 30-year-old who did not.

Died, Age 30

P(Survived) = .4077

Survived, Age 20

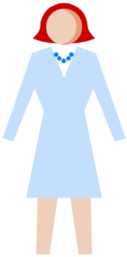
P(Survived) = .4272

The actual sorting agrees with the model.
This is a **concordant** pair.

Discordant Pair

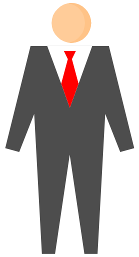
- Compare a 45-year-old who survived with a 35-year-old who did not.

Died, Age 35



P(Survived) = .3981

Survived, Age 45



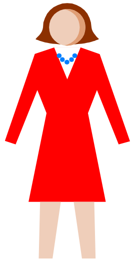
P(Survived) = .3791

The actual sorting disagrees with the model.
This is a **discordant** pair.

Tied Pair

- Compare two 50-year-olds. One survived and the other did not.

Died, Age 50



P(Survived) = .3697

Survived, Age 50

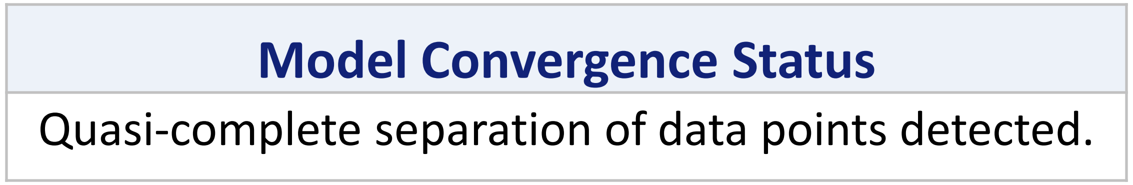
P(Survived) = .3697

The model cannot distinguish between the two.
This is a **tied** pair.

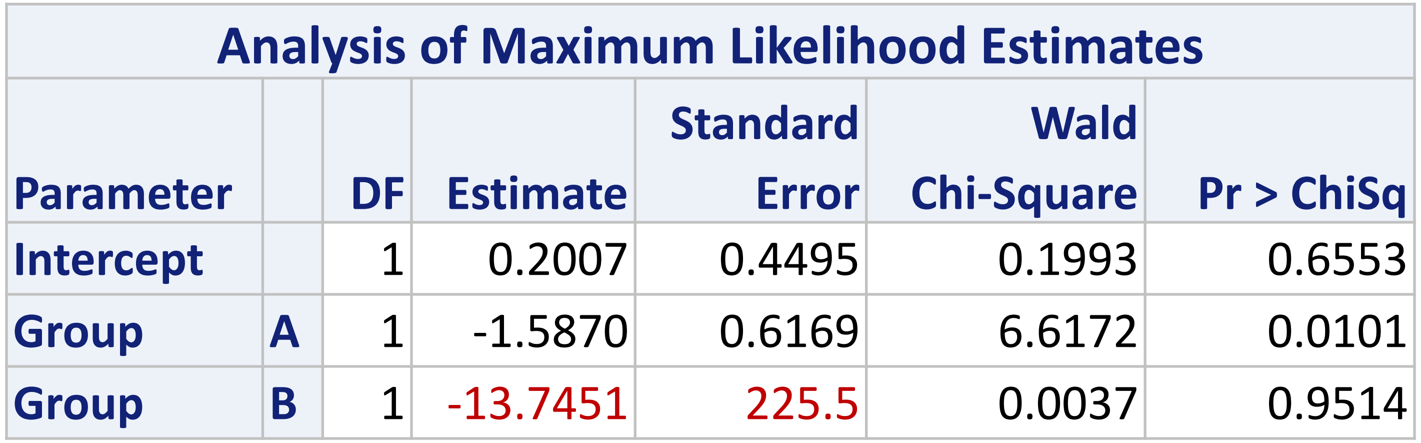
Model: Concordant, Discordant, and Tied Pairs

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	51.3	Somers' D	0.050
Percent Discordant	46.4	Gamma	0.051
Percent Tied	2.3	Tau-a	0.024
Pairs	264313	c	0.525

Quasi-Complete Separation

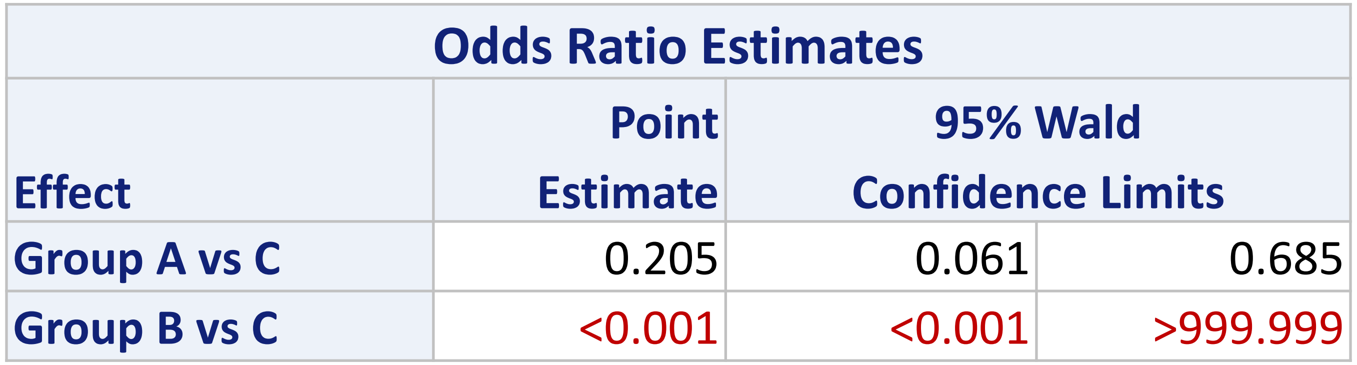


Model Convergence Status
Quasi-complete separation of data points detected.



Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.2007	0.4495	0.1993	0.6553
Group	A	1	-1.5870	0.6169	6.6172	0.0101
Group	B	1	-13.7451	225.5	0.0037	0.9514



Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Group A vs C	0.205	0.061	0.685
Group B vs C	<0.001	<0.001	>999.999

Quasi-Complete Separation

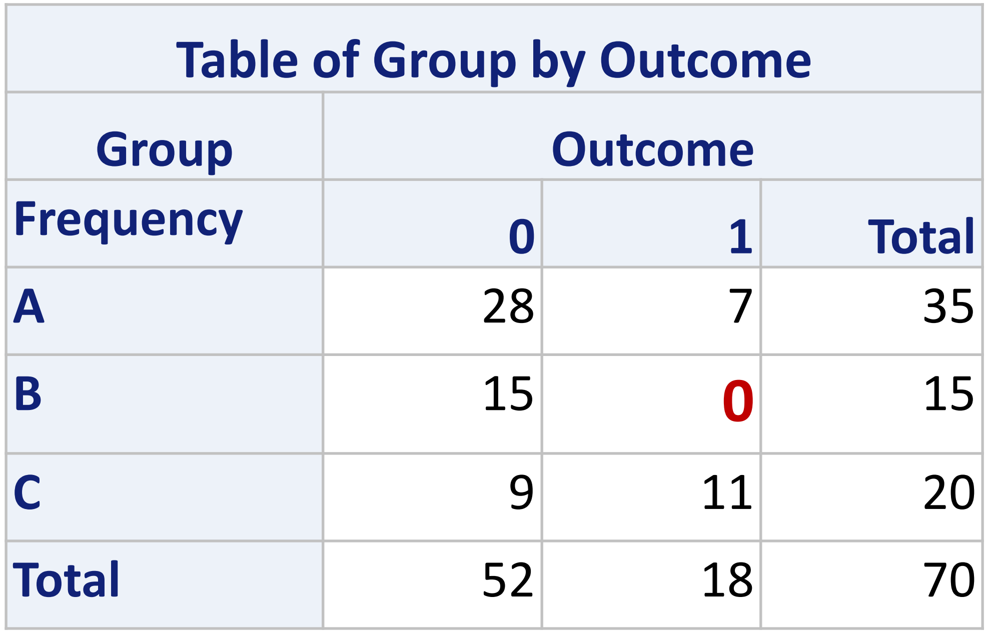


Table of Group by Outcome

Group	Outcome		
Frequency	0	1	Total
A	28	7	35
B	15	0	15
C	9	11	20
Total	52	18	70

LOGISTIC Procedure

```
PROC LOGISTIC <options>;
     CLASS variable</v-options>;
     MODEL response = <effects></options>;
     CONTRAST 'label' effect values</options>;
     EXACT <'label'><Intercept><effects></options>;
     ODDSRATIO <'label'> variable </ options>;
     ROC <'label'> <specification> </ options>;
     ROCCONTRAST <'label'><contrast></ options>;
     SCORE <options>;
     STRATA effects</options>;
     UNITS predictor1=list1 </option>;
     OUTPUT <OUT=SAS-data-set> keyword=name…
                  keyword=name></option>;
RUN;
```

Fitting Simple Binary Logistic Regression Models

This demonstration illustrates the concepts discussed previously.

Demo – Viya – Building Models

https://www.sas.com/en_us/learn/academic-programs.html

Questions?

tom.grant@sas.com