Video 13 - Statistical models

Steve Simon

Measurement

- Traditional levels (scales) of measurement
  - Nominal
  - Ordinal
  - Interval
  - Ratio
- Special cases
  - Binary data
  - Count data, rate data
  - Time-to-event

Ordinal verus interval controversy

- Sums of ordinal variables are meaningless
- Counterexample: grade point average
  - Shift from A to B versus a shift from D to F?
  - Two B's equal and A plus a C?
- Purist versus pragmatist
- Is a sum of Likert scale items different?
  - Unequal scalings average out?

Permissible statistical summaries

- Nominal: percentage, mode
- Ordinal: median
- Interval: mean, standard deviation
- Ratio: Coefficient of variation
- Special cases

Permissible models

- Special cases
  - Binary: Logistic regression
  - Counts: Poisson regression
  - Time-to-event data: Cox proportional hazards regression
- Nominal: Chi-square tests, multinomial logistic regression
- Ordinal outcome variable: Non-parametric tests, ordinal logistic regression
- Ordinal indepdent variable" p for trend tests
- Interval/ratio: t-tests, analysis of variance, linear regression

First break

- What have you learned?
  - Scales of measurement
  - Ordinal verus interval controversy
- What's coming next?
  - Descriptive statistics
  - Linear regression

Steps in your data analysis

– Quality check of data
– Description of sample
– Test of hypotheses/research questions
– Additional exploratory analyses

Quality check of your data (1/2)

– Completeness of data collection
– Review for responses that are ambiguous, out of range, etc
– Edit responses as needed
– Check response frequencies
  • Combine smaller categories, if needed

Quality check of your data (2/2)

– Zero (or near-zero) variation
– Missing value count
– List five five rows, last five rows
– Correlations

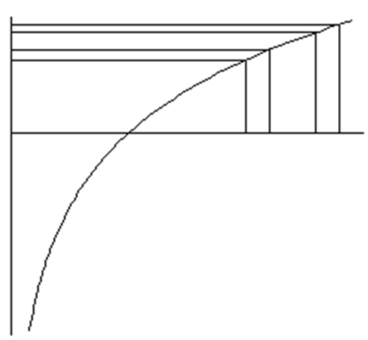
Data reduction

– Check composite scores
  - Check Cronbach's alpha
  - Examine leaving out single items
  - Factor analysis, Structural Equations Modeling

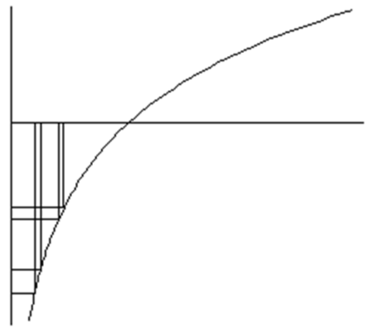
Data transformations

- Ideal - selected a priori
  - Sometimes based on precedent
  - Sometimes motivated by theory
  - Sometimes based on empirical findings
- Don't bother if your range is narrow
  - max/min <= 3
- Log transformation

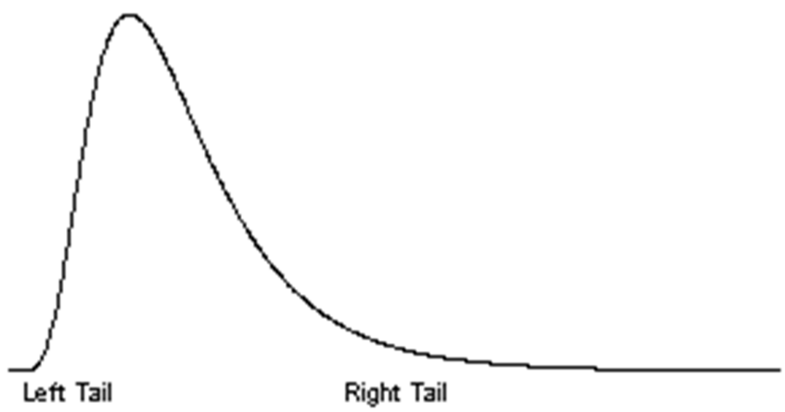
Log transformation



Log transformation



Log transformation



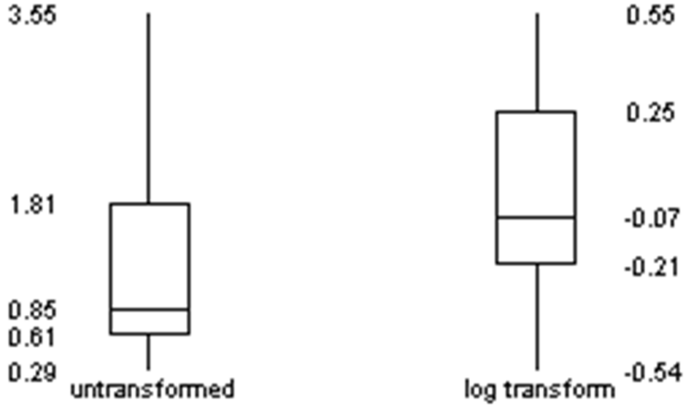
Log transformation fixes

- Skewness
- Outliers
- Unequal variation
- Multiplicative models
  - log(ab) = log(a)+log(b)

When should you use the log transformation?

- Data bounded below by zero.
  - Mean < Standard deviation
- Ratio data
- Max > 3*Min

Log transformation



Log transformation

Standard deviations, untransformed

Report

DM/DX ratio

Functional alleles	Mean	N	Std. Deviation
No functional alleles	1.272	15	1.036
One or more functional alleles	.013	191	.025
Total	.104	206	.426

Standard deviations, log transformed

Report

log DM/DX ratio

Functional alleles	Mean	N	Std. Deviation
No functional alleles	-.018	15	.335
One or more functional alleles	-2.281	191	.531
Total	-2.116	206	.785

Log transformation, summary

- Removes skewness
- Removes outliers
- Stabilizes variances
- Does not always work
- Best when
  - Data bounded below by zero
  - Mean < Standard deviation
  - Max/Min > 3

Descriptive statistics

- Part of every quantitative study
- Table 1, overall summaries
  - Outcomes and covariates
  - Means and standard deviations
  - Percentages (always show denominator)
- Key subgroup comparisons
  - Crosstabulations
  - Means/standard deviations by subgroup

Rules for crosstabulations

– Never display multiple statistics
– Place treatment/exposure categories in the rows
– Summarize using row percentages
– Many rows, not many columns
– Round liberally.

Table of percentages

Table of counts

	Happy	Miserable	Total
Rich	30	10	40
Poor	90	70	160
Total	120	80	200

Table of column percentages

Table of column percents

	Happy	Miserable	Total
Rich	25%	12%	20%
Poor	75%	88%	80%
Total	100%	100%	100%

Table of row percentages

Table of row percents

	Happy	Miserable	Total
Rich	75%	25%	100%
Poor	56%	44%	100%
Total	60%	40%	100%

Table of cell percentages

Table of cell percents

	Happy	Miserable	Total
Rich	15%	5%	20%
Poor	45%	35%	80%
Total	60%	40%	100%

Combining two numbers

Table of counts and row percents

	Happy	Miserable	Total
Rich	75% (30)	25% (10)	100% (40)
Poor	56% (90)	44% (70)	100% (160)
Total	60% (120)	40% (80)	100% (200)

Table of percentages

Alternate display of cell percents

Poor and happy	45%
Poor and miserable	35%
Rich and happy	15%
Rich and miserable	5%

Table of percentages

Table with many rows

	Rich	Poor
Cloud nine	30% (14)	70% (32)
Cheerful	27% (11)	73% (30)
Content	20% (7)	80% (28)
Despondent	16% (5)	84% (26)
Dejected	11% (3)	89% (24)
Depressed	9% (2)	91% (20)
Total	25% (40)	75% (160)

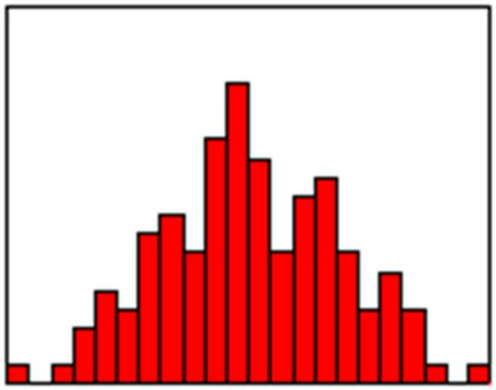
Rules for crosstabulations

– Never display multiple statistics
– Place treatment/exposure categories in the rows
– Summarize using row percentages
– Many rows, not many columns
– Round liberally.

Graphs

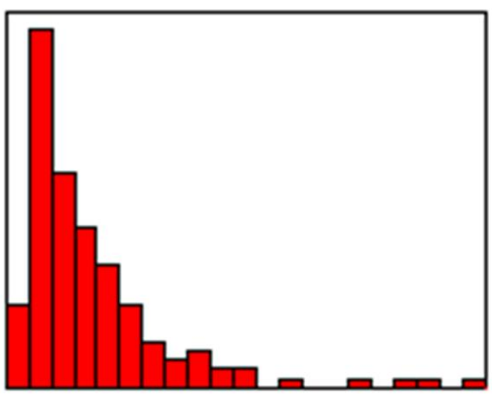
– Overall summaries
  - Histograms for continuous data
  - Bar/pie charts for categorical data
– Assessing relationships
  - Side by side pie/bar charts
  - Boxplots
  - Scatterplots

Histogram examples (1 of 3)



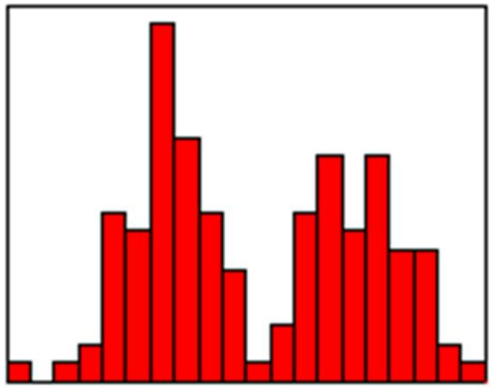
Histogram showing a roughly bell shaped curve

Histogram examples (2 of 3)



Histogram showing a skewed right distribution

Histogram examples (3 of 3)

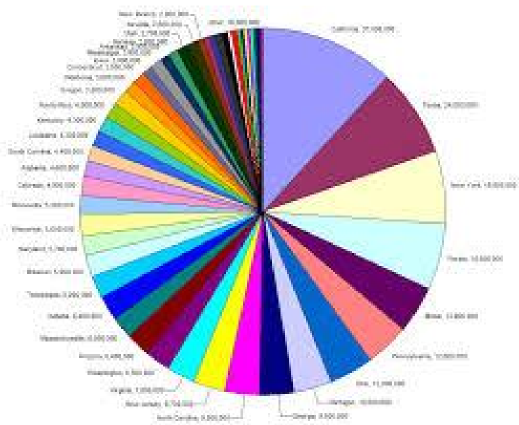


Histogram showing a bimodal distribution

Side by side pie/bar charts

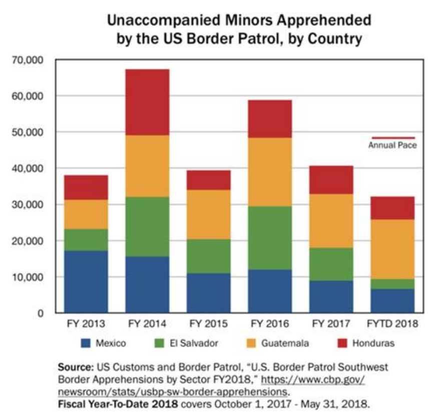
- Pies and bars only work well for 2 or 3 categories
  - Pacman charts
- No good graphs for more categories
- Avoid cheap 3D effects

A very busy pie chart



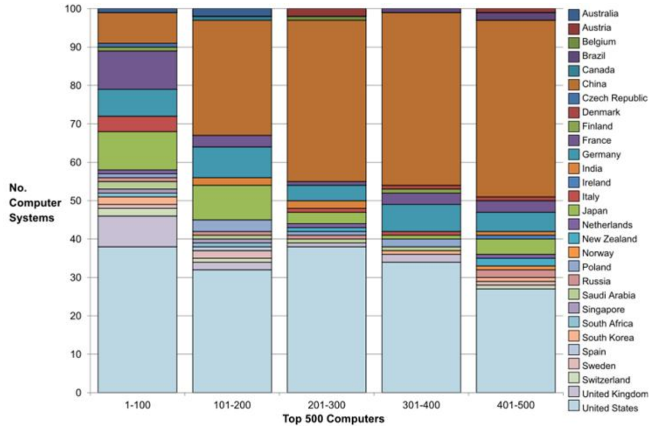
Pie chart of populations for all fifty states

A bar chart



Bar chart showing country of orgin for unaccompanied minors

A very busy bar chart



Bar chart showing distribution of computers across countries

Boxplot

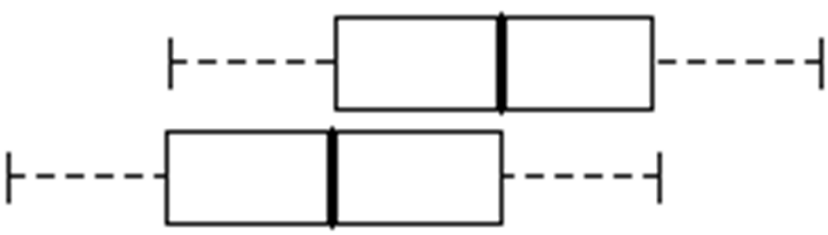
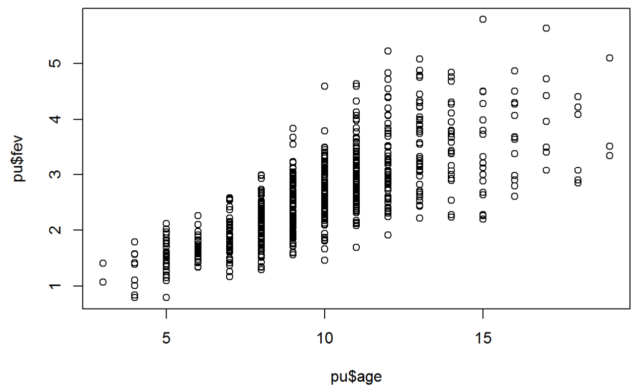


Image of two boxplots

Scatterplot



Second break

– What have you learned?
  • Descriptive statistics
– What's coming next?
  • Linear regression

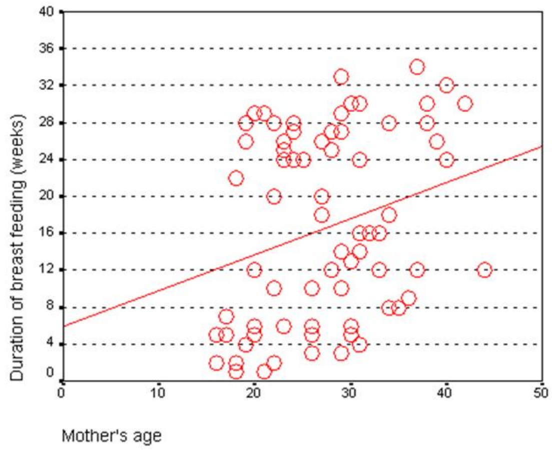
Linear regression

- Continuous outcome variable
- Very flexible
  - Either categorical or continuous independent variables
  - Multiple variables (risk adjustment)
  - Interactions
- Alternatives
  - t-test
  - Analysis of variance

Linear regression

- High school algebra
  - $Y = m X + b$
  - $m = \Delta y / \Delta x$
- The slope represents the estimated average change in Y when X increases by one unit.
- The intercept represents the estimated average value of Y when X equals zero.

Age vs duration - graph



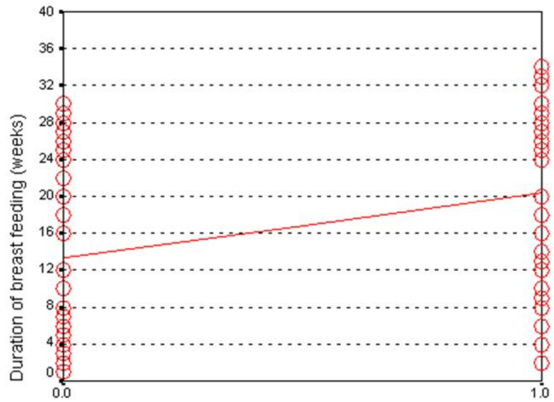
Age vs duration - output

Parameter Estimates

Dependent Variable: Duration of breast feeding (weeks)

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Intercept	5.920	4.580	1.292	.200	-3.195	15.035
MOM_AGE	.389	.162	2.399	.019	6.626E-02	.712

Treatment vs duration



Treatment vs duration

Parameter Estimates

Dependent Variable: Duration of breast feeding (weeks)

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Intercept	20.368	1.569	12.983	.000	17.246	23.491
[FEED_TYP=Control]	-7.050	2.142	-3.292	.001	-11.312	-2.788
[FEED_TYP=Treatmen]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

Adjusted model

- Crude model
  - One independent variable
- Adjusted model
  - More than one independent variable
- Interpretation of slope
  - Estimated average change in Y
  - When X1 changes by one unit
  - And X2 is held contant.

Adjusted model

Parameter Estimates

Dependent Variable: Age when bf stopped

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Intercept	12.961	5.146	2.519	.014	2.719	23.203
MOM_AGE	.249	.165	1.510	.135	-7.919E-02	.577
[FEED_TYP=1]	-5.972	2.241	-2.664	.009	-10.434	-1.511
[FEED_TYP=2]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

Some alternatives

- t-test (two sample t-test)
  - Continuous outcome
  - Catregorical independent variable with two levels
- Disadvantages of the t-test
  - No risk adjustment or interactions
- Analysis of variance
  - Continuous outcome
  - Categorical independent variable with three or more levels
  - Can use more than one categorical independent variable
- Analysis of covariance

Continuous outcomes - summary

- Linear regression
  - Continuous outcome
  - Can provide risk adjustments
- Two-sample t-test
- Analysis of variance
- Analysis of covariance

Third break

- What have you learned?
  - Linear regression
- What’s coming next?
  - Logistic regression
  - Poisson regression

Logistic regression

- Binary outcome variable
- Either categorical or continuous independent variables
- Multiple variables (risk adjustment)
- Interactions

A linear model for probability (1/2)

GA	prob BF
28	60%
29	62%
30	64%
31	66%
32	68%
33	70%
34	72%

Table showing a reasonable linear relationship

A linear model for probability (2/2)

GA	prob BF
28	88%
29	91%
30	94%
31	97%
32	100%
33	103%
34	106%

Table showing an unreasonable linear relationship

A multiplicative model for probability

GA	prob BF
28	0.01%
29	0.03%
30	0.09%
31	0.27%
32	0.81%
33	2.43%
34	7.29%

A reasonable multiplicative model for probability

The relationship between odds and probability

- Usually only seen in gambling contexts
- Sometimes ambiguous
  - Odds in favor versus odds against
- Odds = Prob / (1-Prob)
- Prob = Odds / (1+Odds)

A multiplicative odds model

GA	odds BF
28	27 to 1 against (.037)
29	9 to 1 against (.111)
30	3 to 1 against (.333)
31	1 to 1 (1)
32	3 to 1 in favor (3)
33	9 to 1 in favor (9)
34	27 to 1 in favor (27)

A multiplicative model for odds

Linearity on log-odds scale

GA	odds BF	log odds
28	27 to 1 against (.037)	-3.30
29	9 to 1 against (.111)	-2.20
30	3 to 1 against (.333)	-1.10
31	1 to 1 (1)	0.00
32	3 to 1 in favor (3)	1.10
33	9 to 1 in favor (9)	2.20
34	27 to 1 in favor (27)	3.30

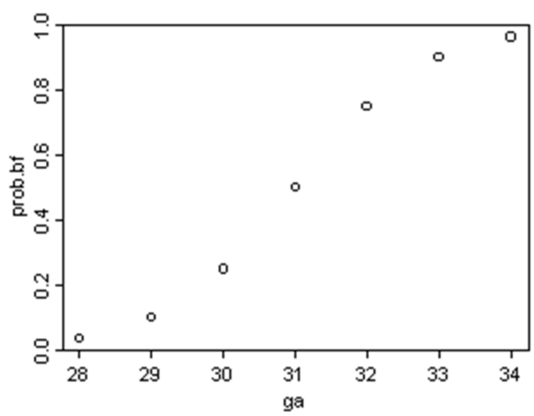
Table showing linearity on the log odds scale

The S-shaped logistic curve (1/2)

GA	odds BF	prob BF
28	27 to 1 against (.037)	3.6%
29	9 to 1 against (.111)	10.0%
30	3 to 1 against (.333)	25.0%
31	1 to 1 (1)	50.0%
32	3 to 1 in favor (3)	75.0%
33	9 to 1 in favor (9)	90.0%
34	27 to 1 in favor (27)	96.4%

Table converting back to probabilities

The S-shaped logistic curve (2/2)



Graph of probabilities

An example of a log odds model with real data (1/2)

GA	Actual prob BF
28	2/6 = 33.3%
29	2/5 = 40.0%
30	7/9 = 77.8%
31	7/9 = 77.8%
32	16/20 = 80.0%
33	14/15 = 93.3%

Log odds model for real data set

An example of a log odds model with real data (2/2)

GA	Predicted log odds	Predicted odds BF	Predicted prob BF
28	-0.57	0.57	36.2%
29	0.01	1.01	50.3%
30	0.59	1.80	64.3%
31	1.16	3.20	76.2%
32	1.74	5.70	85.1%
33	2.32	10.15	91.0%

Log odds model for real data set

Model computations

- log odds = -16.72 + 0.577*GA
- Example: GA=30, estimated probability = 64.3%
  - log odds = -16.72 + 0.577*30 = 0.59
  - odds = exp(0.59) = 1.80
  - prob = 1.80 / (1+1.80) = 0.643
- GS=31
  - log odds = 1.16, odds = 3.20, prob = 76.2%
- GS=32
  - log odds = 1.74, odds = 5.70, prob = 85.1%
- Constant odds ratio
  - 3.20 / 1.80 = 1.78

Categorical variables in a logistic regression (1/2)

sex * survived Crosstabulation

			survived		Total
			No	Yes	
sex	female	Count	154	308	462
		% within sex	33.3%	66.7%	100.0%
	male	Count	709	142	851
		% within sex	83.3%	16.7%	100.0%
Total		Count	863	450	1313
		% within sex	65.7%	34.3%	100.0%

Crosstabulation of gender and mortality on the Titanic

Categorical variables in a logistic regression (2/2)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	SexMale	-2.301	.135	291.069	1	.000	.100
	Constant	.693	.099	49.327	1	.000	2.000

a. Variable(s) entered on step 1: SexMale.

Logistic output for the Titanic data

Alternatives to logistic regresion

– Test of two proportions
  - Only for a binary independent variable
  - No risk adjustments or interactions
– Chisquare test
  - Only for a categorical independent variable
  - Either two or more than two levels

What if your outcomes is categorical but not binary?

- Three of more levels
  - Chi-square test
  - Multinomial logistic regression
- Ordinal outcome variable
  - Nonparametric tests
  - Ordinal logistic regression

Categorical outcomes – summary

- Logistic regression
  - Binary outcome variable
  - Both categorical and continuous independent variables
  - Risk adjustmentsn and interactions possible
- Alternative methods
  - Test of two proportions
  - Chi-square test
  - Multinomial logistic regression.
  - Nonparametric tests
  - Ordinal logistic regression

Poisson regression

– The problems with counts
  - Skewed
  - Non-negative
  - Unequal variances

– Analysis of rates

Poisson regression example - data

– Responses to a mailing
  - 0 9
  - 1 4
  - 2 2
  - 3 3
  - 4 0
  - 5 0
  - 6 1

Poisson regression example - output

```
Call: glm(formula = ct ~ tm, family = poisson)

Coefficients:
(Intercept)           tm
     2.1063      -0.5505

Degrees of Freedom: 6 Total (i.e. Null); 5
Residual
```

– exp(2.1063) = 8.2
– exp(-0.5505) = 0.58

Poisson regression example: Predictions

```
round(predict(pmod),4)
     1      2      3      4       5       6
7
2.1063 1.5558 1.0053 0.4548 -0.0957 -0.6462 -
1.1967

> round(exp(predict(pmod)),4)
     1      2      3      4      5      6      7
8.2177 4.7388 2.7327 1.5758 0.9087 0.5240 0.3022
```

– 4.7388 / 8.2177 = 0.58
– 2.7327 / 4.7388 = 0.58

Fourth break

- What have you learned?
  - Logistic regression
  - Poisson regression
- What’s coming next?
  - Cox regression
  - Longitudinal/hierarchical designs

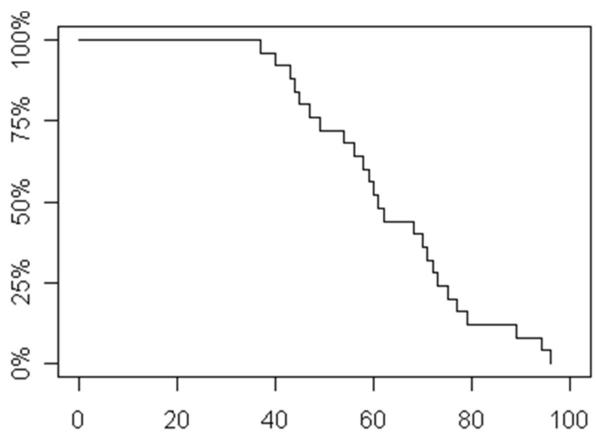
Time to event outcomes

- Special type of ratio scale outcome
  - Non-negative
  - Usually skewed
- Censoring
  - Partial information on some subjects
  - Not the same as missing data

Fruit fly experiment - the data

```
Day Prob Day Prob Day Prob
 37 96%   40 92%   43 88%
 44 84%   45 80%   47 76%
 49 72%   54 68%   56 64%
 58 60%   59 56%   60 52%
 61 48%   62 44%   68 40%
 70 36%   71 32%   72 28%
 73 24%   75 20%   77 16%
 79 12%   89  8%   94  4%
 96  0%
```

Here's a graph of these probabilities over time.

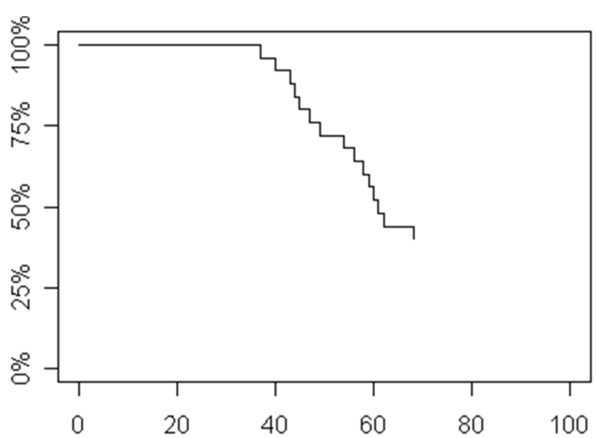


Plot of survival probabilites from first experiment

The data, round 2

```
Day Prob  Day Prob  Day Prob
 37 96%    40 92%    43 88%
 44 84%    45 80%    47 76%
 49 72%    54 68%    56 64%
 58 60%    59 56%    60 52%
 61 48%    62 44%    68 40%
 70+ ?     70+ ?     70+ ?
 70+ ?     70+ ?     70+ ?
 70+ ?     70+ ?     70+ ?
 70+ ?
```

The graph, round 2

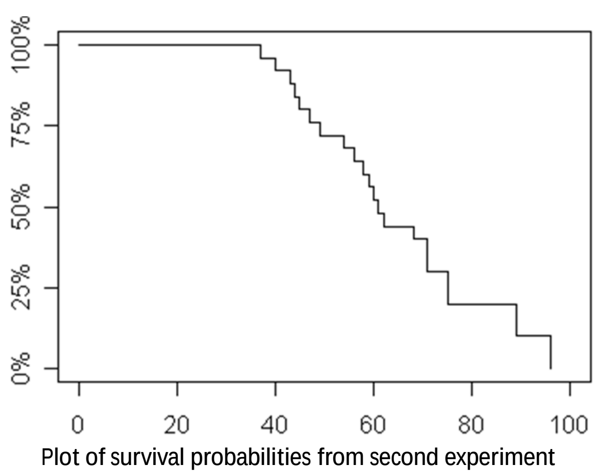


Plot of survival probabilities from second experiment

The data, round 3

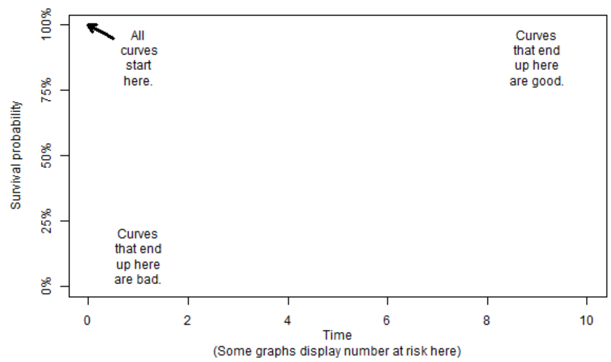
```
Day Prob  Day Prob  Day Prob
 37 96%    40 92%    43 88%
 44 84%    45 80%    47 76%
 49 72%    54 68%    56 64%
 58 60%    59 56%    60 52%
 61 48%    62 44%    68 40%
 70+ ?     71 30%    70+ ?
 70+ ?     75 20%    70+ ?
 70+ ?     89 10%    70+ ?
 96  0%
```

The graph, round 3



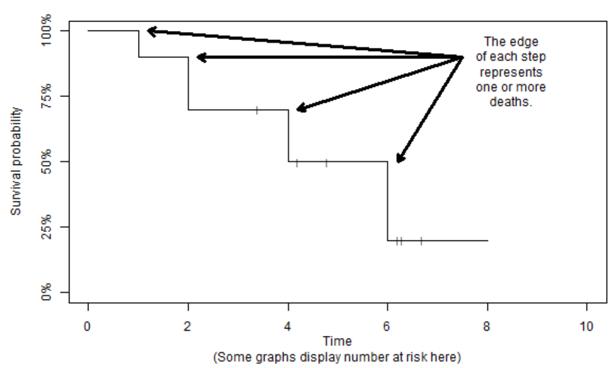
Plot of survival probabilities from second experiment

Happy and sad corners for the Kaplan-Meier curve



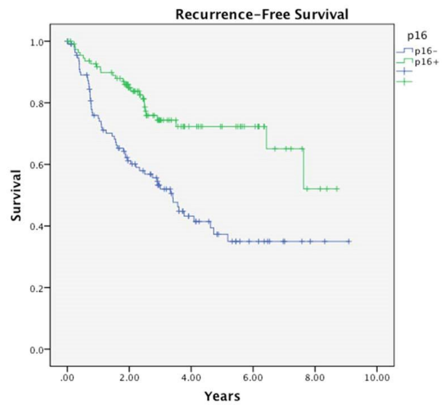
conceptual illustration of Kaplan-Meier curve regions

Cox regression



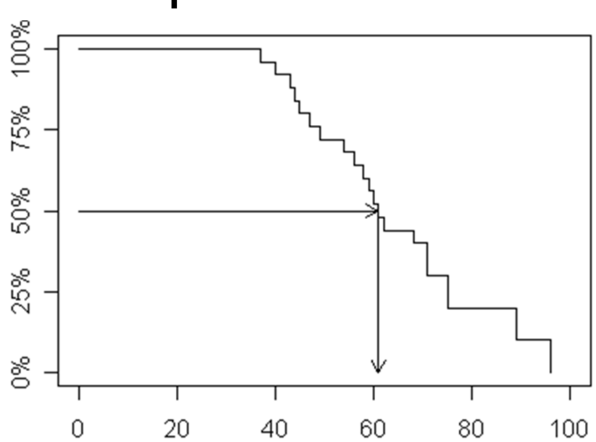
Stair steps in a Kaplan-Meier curve

Cox regression



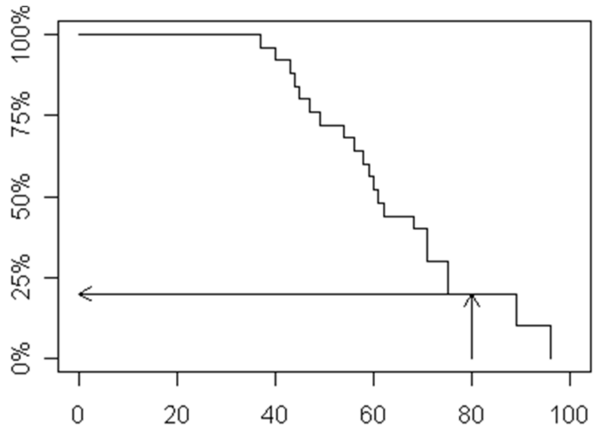
Simple example comparing two Kaplan-Meier curves

Estimating the median and other percentiles



Schematic of survival curve starting from upper left corner

Estimating a fixed time survival probability



Schematic of survival curve starting from lower left corner

Cox regression

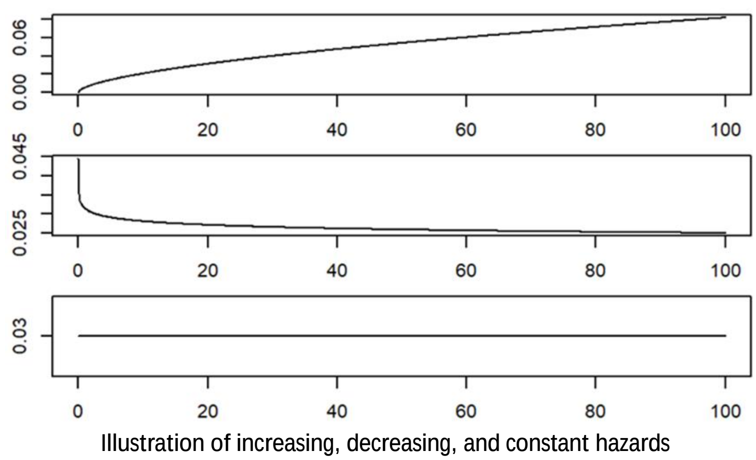


Illustration of increasing, decreasing, and constant hazards

Alternatives to Cox regression

- Log rank test
  - Single categorical indpendent variable
  - Any number of levels
- Parametric survival models
  - Requires much stronger assumptions
  - Exponential, Weibull, or other distribution
  - Can extrapolare beyond the range of the data

Cox regression - summary

- Time-to-event outcome
- Continuous or categorical independent variables
- Mutiple independent variables
  - Risk adjustment
  - Interactions
- Alternatives
  - Log rank test
  - Parametric models

Summary - the big four models

- Linear regression
- Logistic regression
- Poisson regression
- Cox regression
- All very flexible
  - Allow categorical and continuous independent variables
  - Allow for risk adjustments and interactions

Hierarchical/longitudinal designs

- Matching
- Baseline measures
- Longitudinal designs
- Cluster effects

Matching

- Greatly improves precision
- Logistical issues
  - Close but not exact matches
  - Loss of data due to mismatches
  - Best when controls come from a large pool
- Analysis methods
  - Paired t-test
  - Random effects models

Baseline measures

- Nice to have
  - Adjust for baseline imbalance
  - Improve precision
- Analysis methods
  - Change score
  - Baseline covariate
  - Bonate, Analysis of Pretest-Posttest Designs

Longitudinal designs (1/2)

- Advantages
  - Rich, complete picture
  - Improved precision
- Disadvantages
  - Expensive
  - Dropout

Longitudinal designs (2/2)

- Analysis methods
  - Within subject designs
  - Nested effects
  - Repeated measures
  - Split plot designs
  - Random effects models

Cluster effects

- More than one source of variation
- Sources
  - Families
  - Clinics/Hospitals
  - Schools
  - Multicenter trials
- What is the unit of randomization?
- Analysis methods
  - Random effects models
  - Hierarchical models

Fifth break

- What have you learned?
  - Cox regression
  - Longitudinal/hierarchical designs
- What's coming next?
  - Qualitative data analysis

Analysis of Qualitative Data-resources

– Typically, a one-hour interview requires a minimum of three to four hours (or more) of analysis.
– Involve the participants in the process, especially for narrative research.
– Tools:
  - focus groups
  - semi-structured interviews
  - participant observation
  - archival records

Inductive process

– Start with the specific (raw data / transcript)
  - Develop a theoretical framework from the data
  - Conceptual categories emerge from the data
  - Iterative process
– Define the process
  - Who does the work
  - Privacy protections
  - How you will adapt

Analysis process for qualitative data

– Your research question is only your starting point.
– Don't let your question blind you to new information
– Build themes before you complete your data collection
  - Check back against the raw data
  - Look for negative examples
  - Don't ignore infrequently voiced themes
– When have you achieved saturation?

Coding the texts

– Balancing act
  - Level of creativity by coder to identify categories/relationships
  - Must reflect the informants thoughts
  - Audit of the coding by an independent person can check for the match between the coding and the source information
– Look for "negative cases"

What goes in the methods section of a qualitative study

- Recruitment process
- Structure of the interview/focus group
- Recording and transcription details
- Softare used to create categories
- Process to insure reliability
  - Multiple raters
  - Adjudication of disagreement
  - Other audits

Sixth break

- What have you learned?
  - Qualitative data analysis
- What’s coming next?
  - Writing a methods section

What purpose does a methods section serve?

- Assessment of the quality of your research
  - Brag here about your rigor
  - Save limitations for discussion
- Allow others to replicate/extend
  - Non-obvious details

What should not be included in the methods section

- "The Methods section should include only information that was available at the time the plan or protocol for the study was being written; all information obtained during the study belongs in the Results section."
  - Uniform requirements for manuscripts submitted to biomedical journals: Writing and editing for biomedical publication. J Pharmacol Pharmacother. 2010;1(1):42–58.
- Exceptions
  - Patient counts, Dropout rates, Protocol changes

What belongs in the methods section

- Every methods section is different
- General structure
  - Participants
  - Materials
  - Procedures
  - Measures
  - Analysis

Participants

- Where you will find your participants
- Inclusion/exclusion criteria
- Efforts to insure representativeness

Materials/Procedures

- Only document the non-routine
- Materials
  - Chemicals
  - Include company and location
- Procedures
  - Running complex equipment
  - Multiple step laboratory methods

Measures

- Outcome variables
- Independent variables
- Covariates
- Validity/reliability

Analysis

- Research hypotheses / questions
- Sample size justification
- Descriptive methods
  - Boilerplate: “Continuous variables were summarized as means and SDs, and categorical variables were summarized as percentages.” Saleem 2019.

Analysis

- Statistical model
- Adjustments for multiplicity
- Handling missing values/dropout
- Alpha level and one/two sided tests
  - Boilerplate: “All tests were two sided, and P values below the 5% level were regarded as significant.” Lokken 1995.

Conclusion

– Scales of measurement
– Descriptive statistics
– Linear, logistic, Poisson, and Cox regression
– Analysis of qualitative data
– Writing a methods section