Comments for MEDB 5502, Week 06

Topics to be covered

- What you will learn
  - Test of two proportions
  - Chi-square test of independence
  - Odds ratio versus relative risk
  - Concepts behind the logistic regression model
  - Logistic regression with categorical variables
  - Logistic regression with interactions
  - Risk adjustment
  - Diagnostics

Comparing two binary outcomes

- Is there a difference in the proportion of deaths between male passengers and female passengers on the Titanic?
- Is there difference in the proportion of patients finishing the full three doses of HPV vaccine between Black women and White women?
- Does using a ng tube for feeding in pre-term infants increase the probability of successful breast feeding at six months?

Speaker notes

Most of the statistics, you have seen so far involve a continuous outcome. You can, however, use a binary outcome. Here are three examples comparing a binary outcome between two groups.

Other comparisons involving a binary outcome

- Is there are difference in the proportion of deaths between first class, second class, and third class passengers?
- Does age influence the proportion of women finishing the full three doses of HPV vaccine?
- Controlling for the mother's age, does using a ng tube for feeding in pre-term infants increase the probability of successful breast feeding at six months?

Speaker notes

Here are some more complex comparisons involving a binary outcome. The first example involves a comparison of three proportions, not two. The next example involves a continuous predictor of a binary outcome. The final example involves a comparison of binary outcomes in two groups, but controlling for a third variable.

Hypothesis framework

- $H_0: \pi_1 = \pi_2$
- $H_1: \pi_1 = \pi_2$
- Compute $\hat{p}_1$ and $\hat{p}_2$ from samples
- Accept H_0 if $\hat{p}_1 - \hat{p}_2$ is close to zero.
  - $T = (\hat{p}_1 - \hat{p}_2)/s.e.$
  - 95% CI: $(\hat{p}_1 - \hat{p}_2) \pm Z_{\alpha/2} s.e.$

Speaker notes

The hypothesis to test two proportions uses the symbols π_1 and π_2 to represent the proportions in a population.

Titanic data, 1 of 3

```
data_dictionary: titanic.txt
description: Mortality among passengers of the Titanic
```

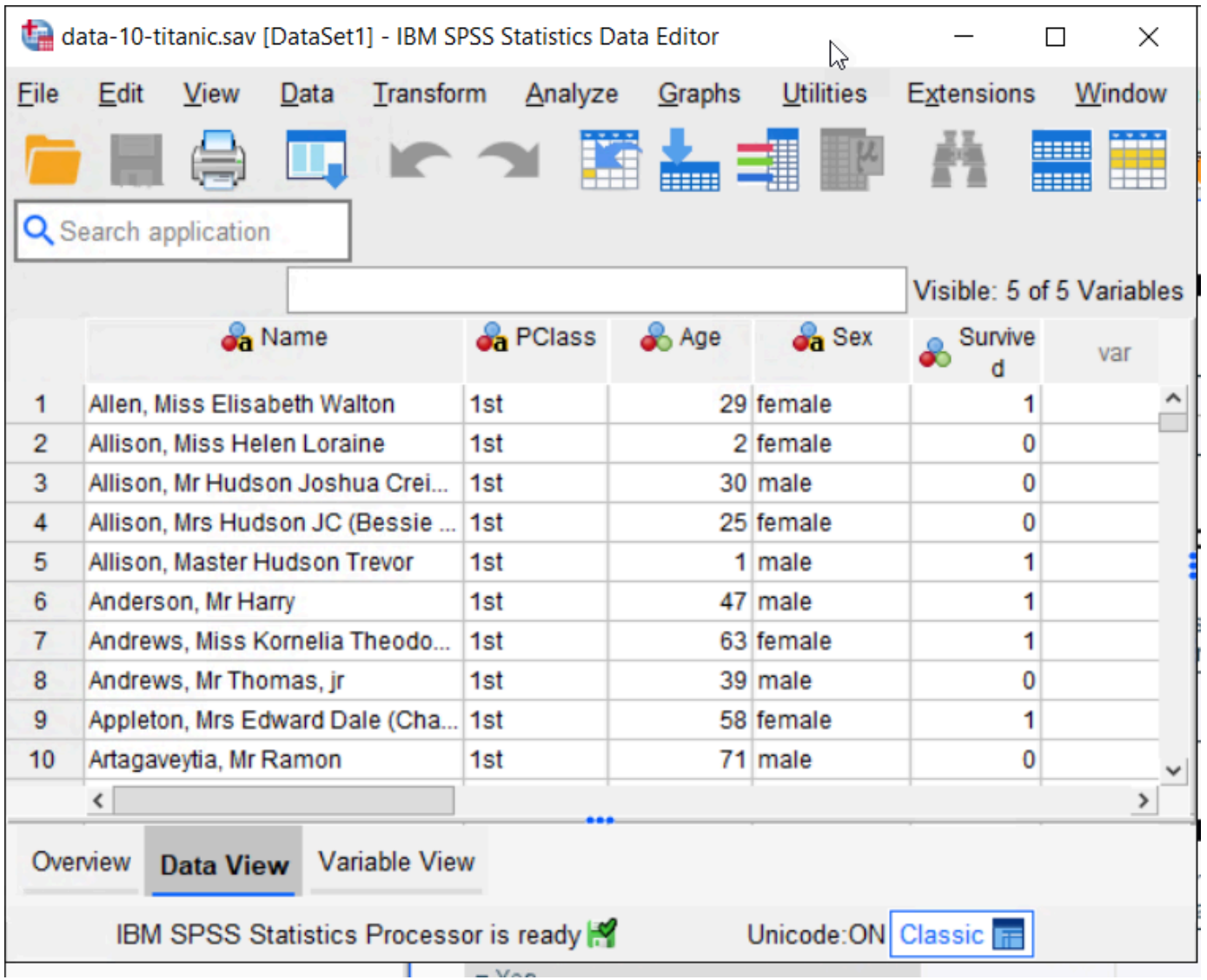
Titanic data, 2 of 3

```
Name:
  label: Passenger name
PClass:
  label: Passenger class
  scale: ordinal
  values: 1st, 2nd, 3rd
Age:
  unit: years
  scale: positive discrete
  missing: NA
```

Titanic data, 3 of 3

```
Sex:
  scale: binary
  values: female, male
Survived:
  scale: binary
  values:
    '1': yes
    '0': no
```

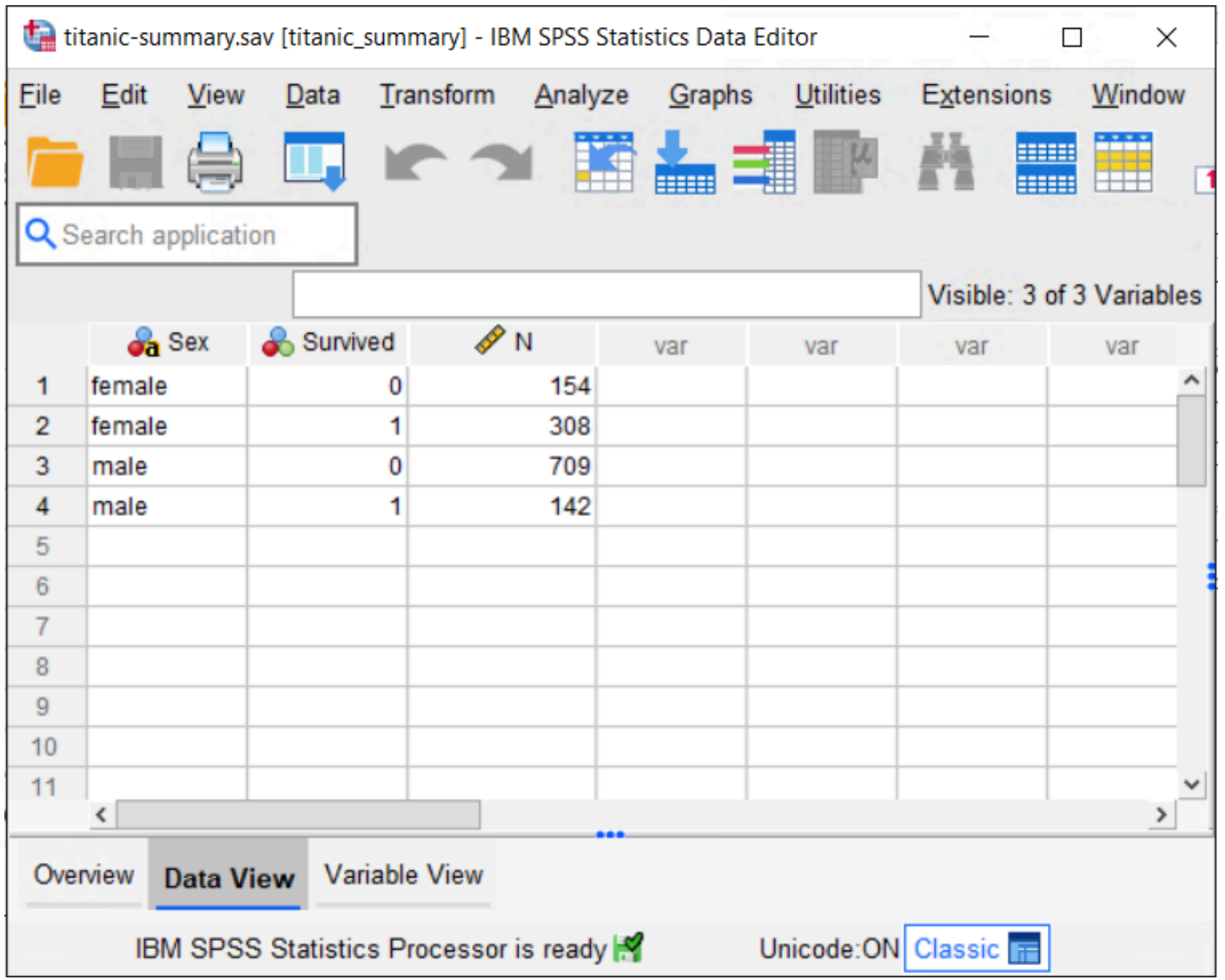
Data layout, 1 of 2



Speaker notes

There are two ways to present the data to SPSS. You can have one row per patient for a total of 1,313 rows.

Data layout, 2 of 2



Speaker notes

Or you could have four rows, one for each combination of sex and survival.

Confidence interval and test of hypothesis

Independent-Samples Proportions Group Statistics

	Sex	Successes	Trials	Proportion	Asymptotic Standard Error
Did the passenger survive? = Yes	= female	308	462	.667	.022
	= male	142	851	.167	.013

Independent-Samples Proportions Confidence Intervals

	Interval Type	Difference in Proportions	Asymptotic Standard Error	95% Confidence Interval of the Difference	
				Lower	Upper
Did the passenger survive? = Yes	Wald	.500	.025	.450	.550

Independent-Samples Proportions Tests

	Test Type	Difference in Proportions	Asymptotic Standard Error	Z	Significance	
					One-Sided p	Two-Sided p
Did the passenger survive? = Yes	Wald H0	.500	.025	18.222	<.001	<.001

Speaker notes

Here is the output from SPSS. The confidence interval contains only positive values, so you can conclude that the difference in proportions is statistically significant.

You can draw the same conclusion from the p-value, which is less than 0.001.

Live demo, Test of two proportions

Break #1

- What you have learned
  - Test of two proportions
- What’s coming next
  - Chi-square test of independence

Chi-square test of independence, 1 of 2

- Equivalent to test of two proportions
- Lay out data in two by two table

	$No\ event$	$Event$
$Treatment$	O_{11}	O_{12}
$Control$	O_{21}	O_{22}

Speaker notes

Chi-square test of independence, 2 of 2

	$No\ event$	$Event$
$Treatment$	$E_{11} = n_1(1 - \hat{p}_{.})$	$E_{12} = n_1\hat{p}_{.}$
$Control$	$E_{21} = n_2(1 - \hat{p}_{.})$	$E_{22} = n_2\hat{p}_{.}$

- $X^2 = \Sigma\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

Speaker notes

Example: Titanic survival by sex

Sex * Did the passenger survive? Crosstabulation

Count

		Did the passenger survive?		
		No	Yes	Total
Sex	female	154	308	462
	male	709	142	851
Total		863	450	1313

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	332.057[a]	1	<.001		
Continuity Correction[b]	329.842	1	<.001		
Likelihood Ratio	332.534	1	<.001		
Fisher's Exact Test				<.001	<.001
N of Valid Cases	1313				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 158.34.

b. Computed only for a 2x2 table

Speaker notes

Here is the output from SPSS. Like most other parts of SPSS, the default is to include four different tests. The tests can differ from one another, but in this case, they all tell the same story. To be honest, this is usually the case.

I recommend using the Person Chi-Square if the sample size is moderate or large and Fisher’s Exact Test if the sample size is small.

What makes it small is the expected count. If any expected count is less than 5, then you should rely on Fisher’s Exact Test.

There is a lot of conflict in the research community about the use of a continuity correction.

Live demo, Chi-square test of independence

Break #2

- What you have learned
  - Chi-square test of independence
- What’s coming next
  - Odds ratio versus relative risk

Titanic data

	Survived	Died	Total
Female	308	154	462
Male	142	709	851
Total	450	863	1,313

Speaker notes

Titanic data, odds of death

	Survived	Died	Total	Odds	
Female	308	154	462	2	to 1 against
Male	142	709	851	4.993	to 1 in favor
Total	450	863	1,313		

Odds ratio = 4.993 / 0.5 = 9.986

Speaker notes

Clearly, a male passenger on the Titanic was more likely to die than a female passenger. But how much more likely? You can compute the odds ratio or the relative risk to answer this question.

The odds ratio compares the relative odds of death in each group. For females, the odds were exactly 2 to 1 against dying (154/308=0.5). For males, the odds were almost 5 to 1 in favor of death (709/142=4.993). The odds ratio is 9.986 (4.993/0.5). There is a ten fold greater odds of death for males than for females.

Titanic data, probability of death

	Survived	Died	Total	Probability
Female	308	154	462	0.3333
Male	142	709	851	0.8331
Total	450	863	1,313	

Relative risk = 0.8331 / 0.3333 = 2.5

Speaker notes

The relative risk (sometimes called the risk ratio) compares the probability of death in each group rather than the odds. For females, the probability of death is 33% (154/462=0.3333). For males, the probability is 83% (709/851=0.8331). The relative risk of death is 2.5 (0.8331/0.3333). There is a 2.5 greater probability of death for males than for females.

There is quite a difference. Both measurements show that men were more likely to die. But the odds ratio implies that men are much worse off than the relative risk. Which number is a fairer comparison?

Which is better

- Relative risk is consistent with how most people think, but
  - Relative risk cannot always be computed
  - Relative risk has an ambiguity

Speaker notes

There are three issues here: The relative risk measures events in a way that is interpretable and consistent with the way people really think. The relative risk, though, cannot always be computed in a research design. Also, the relative risk can sometimes lead to ambiguous and confusing situations. But first, we need to remember that fractions are funny.

Fractions are funny

```
----------  ----------
0.8  (4/5)  1.25 (5/4)
0.75 (3/4)  1.33 (4/3)
0.67 (2/3)  1.50 (3/2)
0.50 (1/2)  2.00 (2/1)
----------  ----------
```

Speaker notes

Suppose you invested money in a stock. On the first day, the value of the stock decreased by 20%. On the second day it increased by 20%. You would think that you have broken even, but that’s not true.

Take the value of the stock and multiply by 0.8 to get the price after the first day. Then multiply by 1.2 to get the price after the second day. The successive multiplications do not cancel out because 0.8 * 1.2 = 0.96. A 20% decrease followed by a 20% increase leaves you slightly worse off.

It turns out that to counteract a 20% decrease, you need a 25% increase. That is because 0.8 and 1.25 are reciprocal. This is easier to see if you express them as simple fractions: 4/5 and 5/4 are reciprocal fractions. Listed here is a table of common reciprocal fractions.

Swapping the numerator and denominator

Speaker notes

Sometimes when we are comparing two groups, we'll put the first group in the numerator and the other in the denominator. Sometimes we will reverse ourselves and put the second group in the numerator.

An odds ratio that you compute with males in the numerator and females in the denominator would be 9.986 or about 10. You could, however, just as easily put the females in the numerator and males in the denominator. Then your odds ratio would be 0.1001 or about 1/10. The two fractions are equivalent. Multiplying the odds by a factor of 10 going from females to males is the multiplying the odds by a factor of 1/10 going from males to females.

This ambiguity also appears for relative risks. The probability of death increases by 2.5 if you put males in the numerator, and the probability of death decreases by a factor of 0.4 if you put females in the numerator. Think of these as reciprocals: 2.5 is 5/2 and 0.4 is 2/5.

The numbers may look quite different but as long as you remember what the reciprocal fraction is, you shouldn't get too confused.

Interpretability, 1 of 3

- Change from 25% probability to 50% probability
- Change from 3 to 1 odds against to even odds
  - RR = 2, OR = 3

Speaker notes

The most commonly cited advantage of the relative risk over the odds ratio is that the former is the more natural interpretation.

The relative risk comes closer to what most people think of when they compare the relative likelihood of events. Suppose there are two groups, one with a 25% chance of mortality and the other with a 50% chance of mortality. Most people would say that the latter group has it twice as bad. But the odds ratio is 3, which seems too big. The latter odds are even (1 to 1) and the former odds are 3 to 1 against.

Interpretability, 2 of 3

- Change from 25% probability to 75% probability
- Change from 3 to 1 odds against to 3 to 1 odds in favor
  - RR = 3, OR = 9

Speaker notes

Even more extreme examples are possible. A change from 25% to 75% mortality represents a relative risk of 3, but an odds ratio of 9.

Interpretability, 3 of 3

- Change from 10% probability to 90% probability
- Change from 9 to 1 odds against to 9 to 1 odds in favor
  - RR = 9, OR = 81

Speaker notes

A change from 10% to 90% mortality represents a relative risk of 9 but an odds ratio of 81.

Designs that rule out the use of the relative risk, 1 of 2

	Cancer cases	Controls	Total
Balding	72	82	154
Hairy	55	57	112
Total	129	139	268

Speaker notes

Some research designs, particularly the case-control design, prevent you from computing a relative risk. A case-control design involves the selection of research subjects on the basis of the outcome measurement rather than on the basis of the exposure.

Consider a case-control study of prostate cancer risk and male pattern balding. The goal of this research was to examine whether men with certain hair patterns were at greater risk of prostate cancer. In that study, roughly equal numbers of prostate cancer patients and controls were selected. Among the cancer patients, 72 out of 129 had either vertex or frontal baldness compared to 82 out of 139 among the controls (see table below).

In this type of study, you can estimate the probability of balding for cancer patients, but you can't calculate the probability of cancer for bald patients. The prevalence of prostate cancer was artificially inflated to almost 50% by the nature of the case-control design.

So you would need additional information or a different type of research design to estimate the relative risk of prostate cancer for patients with different types of male pattern balding.

Designs that rule out the use of the relative risk, 2 of 2

	Heart disease	Healthy	Total
Balding	127 (9.4%)	1,224 (90.6%)	1,351
Hairy	548 (6.7%)	7,611 (93.3%)	8,159
Total	675	8,835	9,510

Speaker notes

Contrast this with data from a cohort study of male physicians (Lotufo et al 2000). In this study of the association between male pattern baldness and coronary heart disease, the researchers could estimate relative risks, since 1,446 physicians had coronary heart disease events during the 11-year follow-up period.

For example, among the 8,159 doctors with hair, 548 (6.7%) developed coronary heart disease during the 11 years of the study. Among the 1,351 doctors with severe vertex balding, 127 (9.4%) developed coronary heart disease (see table below). The relative risk is 1.4 = 9.4% / 6.7%.

You can always calculate and interpret the odds ratio in a case control study. It has a reasonable interpretation as long as the outcome event is rare (Breslow and Day 1980, page 70). The interpretation of the odds ratio in a case-control design is, however, also dependent on how the controls were recruited (Pearce 1993).

Covariate adjustments

	Children	No children	Total
Epilepsy	232 (40%)	354 (60%)	586
Control	79 (72%)	30 (28%)	109
Total	311	384	695

Speaker notes

Another situation which calls for the use of odds ratio is covariate adjustment. It is easy to adjust an odds ratio for confounding variables; the adjustments for a relative risk are much trickier.

In a study on the likelihood of pregnancy among people with epilepsy (Schupf and Ottman 1994), 232 out of 586 males with idiopathic/cryptogenic epilepsy had fathered one or more children. In the control group, the respective counts were 79 out of 109 (see table below).

The simple relative risk is 0.55 and the simple odds ratio is 0.25. Clearly the probability of fathering a child is strongly dependent on a variety of demographic variables, especially age (the issue of marital status was dealt with by a separate analysis). The control group was 8.4 years older on average (43.5 years versus 35.1), showing the need to adjust for this variable. With a multivariate logistic regression model that included age, education, ethnicity and sibship size, the adjusted odds ratio for epilepsy status was 0.36. Although this ratio was closer to 1.0 than the crude odds ratio, it was still highly significant. A comparable adjusted relative risk would be more difficult to compute (although it can be done as in Lotufo et al 2000).

Ambiguous and confusing situations

- One hundred pound sack of potatoes
  - 99% water, 1% potato
  - Weighs 1 pound after completely drying
  - Instead dry until 2% potato
    - How much does it weigh then?

Speaker notes

The relative risk can sometimes produce ambiguous and confusing situations. Part of this is due to the fact that relative measurements are often counter-intuitive. Consider an interesting case of relative comparison that comes from a puzzle on the Car Talk radio show. You have a hundred pound sack of potatoes. Let’s assume that these potatoes are 99% water. That means 99 parts water and 1 part potato. These are soggier potatoes than I am used to seeing, but it makes the problem more interesting.

If you dried out the potatoes completely, they would only weigh one pound. But let’s suppose you only wanted to dry out the potatoes partially, until they were 98% water. How much would they weigh then?

The counter-intuitive answer is 50 pounds. 98% water means 49 parts water and 1 part potato. An alternative way of thinking about the problem is that in order to double the concentration of potato (from 1% to 2%), you have to remove about half of the water.

Relative risks have the same sort of counter-intuitive behavior. A small relative change in the probability of a common event’s occurrence can be associated with a large relative change in the opposite probability (the probability of the event not occurring).

Example: physician recommendations

	No cath	Cath	Total
Male patient	34 (9.4%)	326 (90.6%)	360
Female patient	55 (15.3%)	305 (84.7%)	360
Total	89	631	720

Speaker notes

Consider a recent study on physician recommendations for patients with chest pain (Schulman et al 1999). This study found that when doctors viewed videotape of hypothetical patients, race and sex influenced their recommendations. One of the findings was that doctors were more likely to recommend cardiac catheterization for men than for women. 326 out of 360 (90.6%) doctors viewing the videotape of male hypothetical patients recommended cardiac catheterization, while only 305 out of 360 (84.7%) of the doctors who viewed tapes of female hypothetical patients made this recommendation.

The odds ratio is either 0.57 or 1.74, depending on which group you place in the numerator. The authors reported the odds ratio in the original paper and concluded that physicians make different recommendations for male patients than for female patients.

A critique of this study (Schwarz et al 1999) noted among other things that the odds ratio overstated the effect, and that the relative risk was only 0.93 (reciprocal 1.07). In this study, however, it is not entirely clear that 0.93 is the appropriate risk ratio. Since 0.93 is so much closer to 1 and 0.57, the critics claimed that the odds ratio overstated the tendency for physicians to make different recommendations for male and female patients.

Although the relative change from 90.6% to 84.7% is modest, consider the opposite perspective. The rates for recommending a less aggressive intervention than catheterization was 15.3% for doctors viewing the female patients and 9.4% for doctors viewing the male patients, a relative risk of 1.63 (reciprocal 0.61).

This is the same thing that we just saw in the Car Talk puzzler: a small relative change in the water content implies a large relative change in the potato content. In the physician recommendation study, a small relative change in the probability of a recommendation in favor of catheterization corresponds to a large relative change in the probability of recommending against catheterization.

Thus, for every problem, there are two possible ways to compute relative risk. Sometimes, it is obvious which relative risk is appropriate. For the Titanic passengers, the appropriate risk is for death rather than survival.

Example: Breast feeding study

	Continued bf	Stopped bf	Total
Treatment	19 (37.3%)	32 (62.7%)	51
Control	5 (8.8%)	52 (91.2%)	57
Total	24	84	108

Speaker notes

But what about a breast feeding study. Are we trying to measure how much an intervention increases the probability of breast feeding success or are we trying to see how much the intervention decreases the probability of breast feeding failure? For example, Deeks 1998 expresses concern about an odds ratio calculation in a study aimed at increasing the duration of breast feeding. At three months, 32/51 (63%) of the mothers in the treatment group had stopped breast feeding compared to 52/57 (91%) in the control group.

While the relative risk of 0.69 (reciprocal 1.45) for this data is much less extreme than the odds ratio of 0.16 (reciprocal 6.2), one has to wonder why Deeks chose to compare probabilities of breast feeding failures rather than successes. The rate of successful breast feeding at three months was 4.2 times higher in the treatment group than the control group. This is still not as extreme as the odds ratio; the odds ratio for successful breast feeding is 6.25, which is simply the inverse of the odds ratio for breast feeding failure.

One advantage of the odds ratio is that it is not dependent on whether we focus on the event’s occurrence or its failure to occur. If the odds ratio for an event deviates substantially from 1.0, the odds ratio for the event’s failure to occur will also deviate substantially from 1.0, though in the opposite direction.

Break #3

- What you have learned
  - Odds ratio versus relative risk
- What’s coming next
  - Concepts behind the logistic regression model

What is logistic regression?

- Binary outcome
- Categorical or continuous predictors
- Linear on the log odds scale

Speaker notes

The logistic regression model is a model that uses a binary (two possible values) outcome variable. Examples of a binary variable are mortality (live/dead), and morbidity (healthy/diseased). Sometimes you might take a continuous outcome and convert it into a binary outcome. For example, you might be interested in the length of stay in the hospital for mothers during an unremarkable delivery. A binary outcome might compare mothers who were discharged within 48 hours versus mothers discharged more than 48 hours.

The covariates in a logistic regression model represent variables that might be associated with the outcome variable. Covariates can be either continuous or categorical variables.

For binary outcomes, you might find it helpful to code the variable using indicator variables. An indicator variable equals either zero or one. Use the value of one to represent the presence of a condition and zero to represent absence of that condition. As an example, let 1=diseased, 0=healthy.

Why log odds?

- Statistical model of surgery
  - Estimates probability of demise
  - First prediction: probability=1.2
- Log odds prevent out of range predictions

Speaker notes

A logistic regression model examines the relationship between one or more independent variable and the log odds of your binary outcome variable. Log odds seem like a complex way to describe your data, but when you are dealing with probabilities, this approach leads to the simplest description of your data that is consistent with the rules of probability.

Let’s consider an artificial data example where we collect data on the gestational age of infants (GA), which is a continuous variable, and the probability that these infants will be breast feeding at discharge from the hospital (BF), which is a binary variable. We expect an increasing trend in the probability of BF as GA increases. Premature infants are usually sicker and they have to stay in the hospital longer. Both of these present obstacles to BF.

A linear model for probability, 1 of 2

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

Speaker notes

A linear model would presume that the probability of BF increases as a linear function of GA. You can represent a linear function algebraically as

prob BF = a + b*GA

This means that each unit increase in GA would add b percentage points to the probability of BF. The table shown below gives an example of a linear function.

Figure 1. Hypothetical probabilities from an additive model

This table represents the linear function

prob BF = 4 + 2*GA

which means that you can get the probability of BF by doubling GA and adding 4. So an infant with a gestational age of 30 would have a probability of **4+2*30 = 64**.

A simple interpretation of this model is that each additional week of GA adds an extra 2% to the probability of BF. We could call this an additive probability model.

A linear model of probability, 2 of 2

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

Speaker notes

I’m not an expert on BF; what little experience I’ve had with the topic occurred over 67 years ago. But I do know that an additive probability model tends to have problems when you get probabilities close to 0% or 100%**. Let’s change the linear model slightly to the following:

prob BF = 4 + 3*GA

This model would produce the following table of probabilities.

Figure 2. Hypothetical probabilities from an alternative additive model

You may find it difficult to explain what a probability of 106% means. This is a reason to avoid using a additive model for estimating probabilities. In particular, try to avoid using an additive model unless you have good reason to expect that all of your estimated probabilities will be between 20% and 80%.

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%

Speaker notes

It’s worthwhile to consider a different model here, a multiplicative model for probability, even though it suffers from the same problems as the additive model.

In a multiplicative model, you change the probabilities by multiplying rather than adding. Here’s a simple example.

Figure 3. Hypothetical probabilities from a multiplicative model

In this example, each extra week of GA produces a tripling in the probability of BF. Contrast this to the linear models shown above, where each extra week of GA adds 2% or 3% to the probability of BF.

A multiplicative model can’t produce any probabilities less than 0%, but it’s pretty easy to get a probability bigger than 100%. A multiplicative model for probability is actually quite attractive, as long as you have good reason to expect that all of the probabilities are small, say less than 20%.

The relationship between odds and probability

- odds = prob / (1-prob)
- prob = odds / (1+odds)
  - $0 \leq$ prob ≤ 1
  - $0 \leq$ odds $\leq \infty$
    - $0 \leq$ odds against ≤ 1
    - $1 \leq$ odds in favor $\leq \infty$

Speaker notes

Another approach is to try to model the odds rather than the probability of BF. You see odds mentioned quite frequently in gambling contexts. If the odds are three to one in favor of your favorite football team, that means you would expect a win to occur about three times as often as a loss. If the odds are four to one against your team, you would expect a loss to occur about four times as often as a win.

You need to be careful with odds. Sometimes the odds represent the odds in favor of winning and sometimes they represent the odds against winning. Usually it is pretty clear from the context. When you are told that your odds of winning the lottery are a million to one, you know that this means that you would expect to having a losing ticket about a million times more often than you would expect to hit the jackpot.

It's easy to convert odds into probabilities and vice versa. With odds of three to one in favor, you would expect to see roughly three wins and only one loss out of every four attempts. In other words, your probability for winning is 0.75.

If you expect the probability of winning to be 20%, you would expect to see roughly one win and four losses out of every five attempts. In other words, your odds are 4 to 1 against.

The formulas for conversion are

odds = prob / (1-prob)

and

prob = odds / (1+odds).

In medicine and epidemiology, when an event is less likely to happen and more likely not to happen, we represent the odds as a value less than one. So odds of four to one against an event would be represented by the fraction 1/5 or 0.2. When an event is more likely to happen than not, we represent the odds as a value greater than one. So odds of three to one in favor of an event would be represented simply as an odds of 3. With this convention, odds are bounded below by zero, but have no upper bound.

A log odds model for probability, 1 of 4

Speaker notes

Let’s consider a multiplicative model for the odds (not the probability) of BF.

Figure 4. Hypothetical odds from a multiplicative model

This model implies that each additional week of GA triples the odds of BF. A multiplicative model for odds is nice because it can’t produce any meaningless estimates.

A log odds model for probability, 2 of 4

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

Speaker notes

It’s interesting to look at how the logarithm of the odds behave.

Notice that an extra week of GA adds 1.1 units to the log odds. So you can describe this model as linear (additive) in the log odds. When you run a logistic regression model in SPSS or other statistical software, it uses a model just like this, a model that is linear on the log odds scale. This may not seem too important now, but when you look at the output, you need to remember that SPSS presents all of the results in terms of log odds. If you want to see results in terms of probabilities instead of logs, you have to transform your results.

A log odds model for probability, 3 of 4

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%

Speaker notes

Let’s look at how the probabilities behave in this model.

Notice that even when the odds get as large as 27 to 1, the probability still stays below 100%. Also notice that the probabilities change in neither an additive nor a multiplicative fashion.

A log odds model for probability, 4 of 4

Speaker notes

A graph shows what is happening.

The probabilities follow an S-shaped curve that is characteristic of all logistic regression models. The curve levels off at zero on one side and at one on the other side. This curve ensures that the estimated probabilities are always between 0% and 100%.

An example of a log odds model with real data, 1 of 3

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%

Speaker notes

There are other approaches that also work well for this type of data, such as a probit model, that I won’t discuss here. But I did want to show you what the data relating GA and BF really looks like.

An example of a log odds model with real data, 2 of 3

Speaker notes

I’ve simplified this data set by removing some of the extreme gestational ages.

The table below shows the predicted log odds, and the calculations needed to transform this estimate back into predicted probabilities.

An example of a log odds model with real data, 3 of 3

- log odds = -16.72 + 0.577 $\times$ 30 = 0.59
- odds = exp(log odds) = 1.8
- prob = odds / (1+odds) = 0.64

Speaker notes

Let’s examine these calculations for GA = 30. The predicted log odds would be the intercept plus the slope times 30.

Convert from log odds to odds by exponentiating.

And finally, convert from odds back into probability.

prob = 1.80 / (1+1.80) = 0.643

The predicted probability of 64.3% is reasonably close to the true probability (77.8%).

You might also want to take note of the predicted odds. Notice that the ratio of any odds to the odds in the next row is 1.78. For example,

3.20/1.80 = 1.78

5.70/3.20 = 1.78

It’s not a coincidence that you get the same value when you exponentiate the slope term in the log odds equation.

exp(0.59) = 1.78

This is a general property of the logistic model. The slope term in a logistic regression model represents the log of the odds ratio. This represents the increase (decrease) in risk as the independent variable increases by one unit.

Live demo, Concepts behind the logistic regression model

Break #4

- What you have learned
  - Concepts behind the logistic regression model
- What’s coming next
  - Logistic regression with categorical variables

Categorical variables in a logistic regression model, 1 of 3

Speaker notes

You treat categorical variables in much the same way as you would in a linear regression model. Create indicator variables for each level of your categorical variable and then include all but one of them in your model. The category associated with the omitted variable is the reference category.

How would SPSS handle a variable like Passenger Class, which has three levels

1st, 2nd, 3rd?

Here’s a crosstabulation of survival versus passenger class.

Notice that the odds of dying are 0.67 to 1 in 1st class, 1.35 to 1 in 2nd class, and 4.15 to 1 in 3rd class. These are odds in favor of dying. The odds against dying are 1.50 to 1, 0.74 to 1, and 0.24 to 1, respectively.

Categorical variables in a logistic regression model, 2 of 3

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	pclass			159.120	2	.000	
	pclass(1)	1.827	.148	152.162	1	.000	6.212
	pclass(2)	1.121	.154	53.267	1	.000	3.069
	Constant	-1.424	.095	225.403	1	.000	.241

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

- 1.50 / 0.24 = 6.212
- 0.74 / 0.24 = 3.069

Speaker notes

The odds ratio for the pclass(1) row is 6.212, which is equal to 1.50 / 0.24. You should interpret this as the odds against dying are 6 times better in first class compared to third class. The odds ratio for the pclass(2) row is 3.069, which equals 0.74 / 0.24. This tells you that the odds against dying are about 3 times better in second class compared to third class. The Constant row tells you that the odds are 0.241 to 1 in third class.

Categorical variables in a logistic regression model, 3 of 3

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	pclass			159.120	2	.000	
	pclass(1)	-.705	.166	18.050	1	.000	.494
	pclass(2)	-1.827	.148	152.162	1	.000	.161
	Constant	.403	.114	12.550	1	.000	1.496

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

- 0.74 / 1.50 = 0.494
- 0.24 / 1.50 = 0.161

Speaker notes

If you prefer to do the analysis with each of the other classes being compared back to first class, then select FIRST for reference category.

This produces the following output:

Here the pclass(1) row provides an odds ratio of 0.494 which equals 0.74 / 1.50. The odds against dying are about half in second class versus first class. The pclass(2) provides an odds ratio of 0.161 (approximately 1/6) which equals 0.24 / 1.50. The odds against dying are 1/6 in third class compared to first class. The Constant row provides an odds of 1.496 to 1 against dying for first class.

Notice that the numbers in parentheses (pclass(1) and pclass(2)) do not necessarily correspond to first and second classes. It depends on how SPSS chooses the indicator variables. How did I know how to interpret the indicator variables and the odds ratios? I wouldn't have known how to do this if I hadn't computed a crosstabulation earlier. It is very important to do a few simple crosstabulations before you run a logistic regression model, because it helps you orient yourself to the data.

Live demo, Logistic regression with categorical variables

Break #5

- What you have learned
  - Logistic regression with categorical variables
- What’s coming next
  - Logistic regression with interactions

Interactions in logistic regression

- Odds ratios vary by a third factor
- Interpretation is more tedious

Odds ratios for first class

Sex * Survived Crosstabulation[a]

Count

		Survived		
		Died	Survived	Total
Sex	female	9	134	143
	male	120	59	179
Total		129	193	322

a. PClass = 1st

Risk Estimate[a]

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for sex_reversed (male / female)	30.282	14.399	63.685
For cohort Survived = Died	10.652	5.613	20.215
For cohort Survived = Survived	.352	.284	.435
N of Valid Cases	322		

a. PClass = 1st

Odds ratio for second class

Sex * Survived Crosstabulation[a]

Count

		Survived		
		Died	Survived	Total
Sex	female	13	94	107
	male	148	25	173
Total		161	119	280

a. PClass = 2nd

Risk Estimate[a]

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for sex_reversed (male / female)	42.806	20.871	87.794
For cohort Survived = Died	7.041	4.215	11.763
For cohort Survived = Survived	.164	.114	.238
N of Valid Cases	280		

a. PClass = 2nd

Odds ratio for third class

Sex * Survived Crosstabulation[a]

Count

		Survived		
		Died	Survived	Total
Sex	female	132	80	212
	male	441	58	499
Total		573	138	711

a. PClass = 3rd

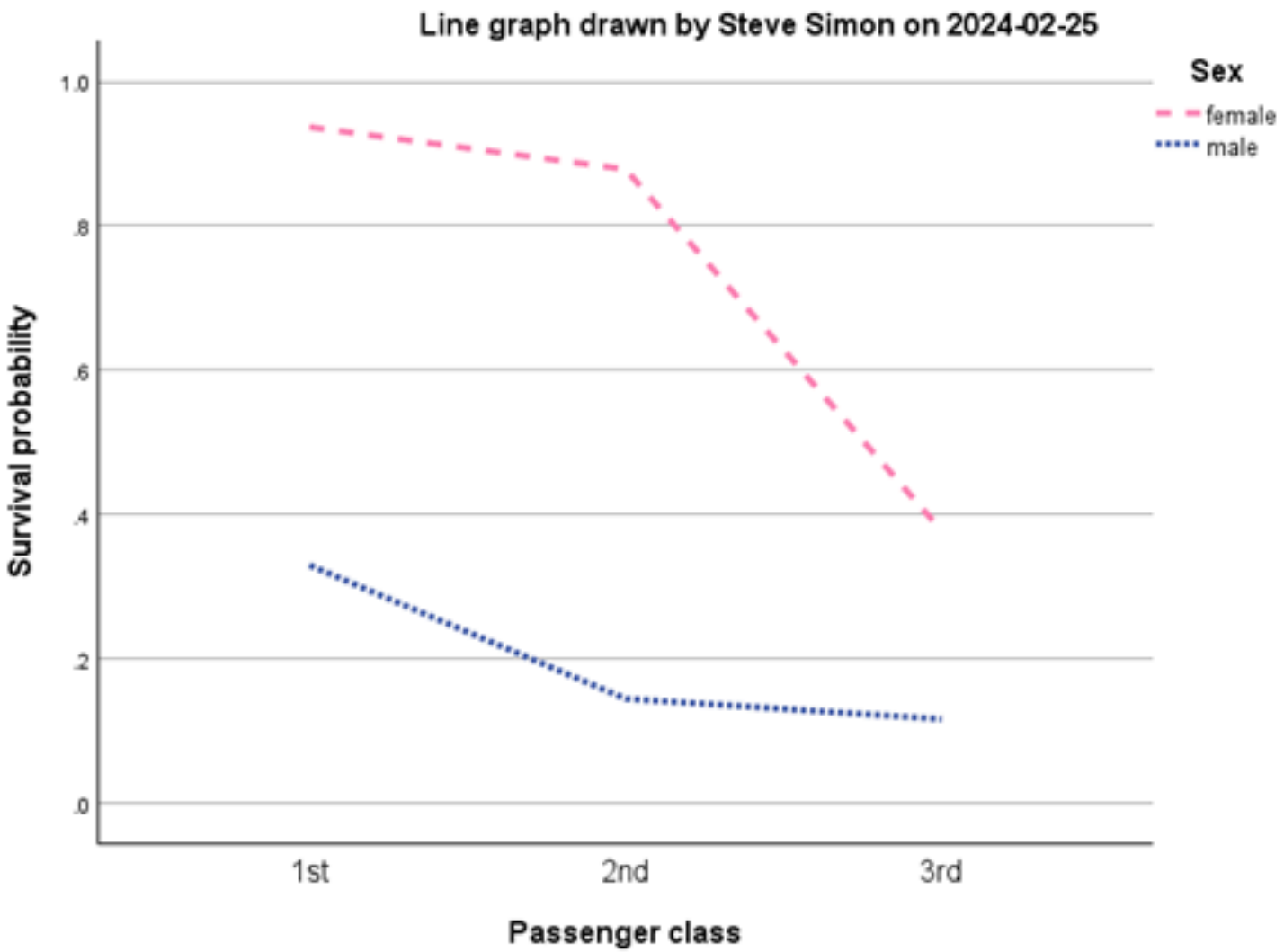
Risk Estimate[a]

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for sex_reversed (male / female)	4.608	3.120	6.806
For cohort Survived = Died	1.419	1.272	1.584
For cohort Survived = Survived	.308	.229	.415
N of Valid Cases	711		

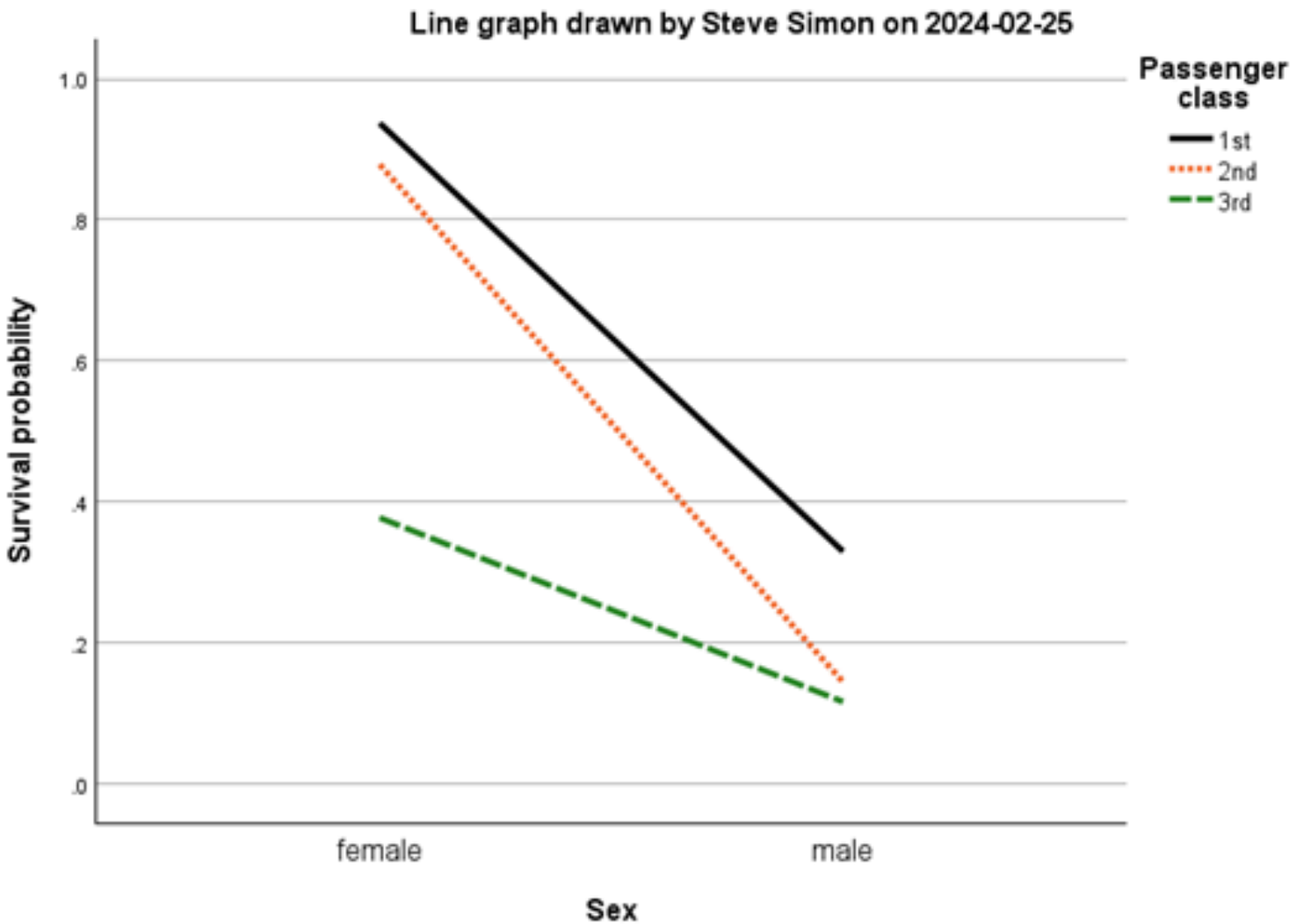
a. PClass = 3rd

Logistic regression with interaction

Line plot for interaction, 1 of 2



Line plot for interaction, 2 of 2



Live demo, Logistic regression with interactions

Break #6

- What you have learned
  - Logistic regression with interactions
- What’s coming next
  - Risk adjustment

Description of bf data, 1 of 11

```
data_dictionary: bf.csv
description: |
  This data comes from a research study done at Children's Mercy Hospital and St. Luke's Medical Center. This was a study of breast feeding in  pre-term infants. Infants were randomized into either a treatment group (NG tube) or a control group (Bottle). Infants in the NG tube group  were fed in the hospital via their nasogastral tube when the mother was not available for breast feeding. Infants in the bottle group received bottles when the mothers were not available. Both groups were monitored for six months after discharge from the hospital.
```

Description of bf data, 2 of 11

```
feed_typ:
  value: Control Treatment
age_stop:
  label: Age at which infant stopped breast feeding
  scale: non-negative real
  unit: weeks
sepsis:
  label: Diagnosis of sepsis
  value: No Yes
total_ab:
  label: Total number of apnea and bradycardia incidents
  scale: non-negative integer
```

Description of bf data, 3 of 11

```
del_type:
  label: Type of delivery
  values:
    Vaginal: 1
    C-section: 2
mom_age:
  label: Mother's age
  unit: years
gravida:
  label: Gravidity or number of pregnancies
  scale: non-negative integer
para:
  label: Parity or number of live births
  scale: non-negative integer
```

Description of bf data, 4 of 11

```
mar_st:
  label: Marital status of mother
  values:
    Single: 1
    Married: 2
race:
  label: Mother's race
  values:
    White: W
    Black: B
smoker:
  label: Smoking by mother during pregnancy
  values:
    'TRUE': 1
    'FALSE': 2
```

Description of bf data, 5 of 11

```
mi_hosp:
  label: Distance from the mother's home to the hospital
  unit: miles
  scale: non-negative integer
ng_tube:
  label: Time on the NG tube
  unit: days
  scale: non-negative integer
tot_bott:
  label: Bottles of formula given while in the hospital
  scale: non-negative integer
```

Description of bf data, 6 of 11

```
bw:
  label: Birthweight
  unit: kg
  scale: non-negative real
gest_age:
  label: Estimated gestational age
  unit: weeks
  scale: positive integer
apgar1:
  label: Apgar score at one minute
  scale: 0 through 10
apgar5:
  label: Apgar score at five minutes
  scale: 0 through 10
```

Description of bf data, 7 of 11

```
bf1_wt:
  label: Weight at first breast feeding
  unit: kg
  scale: non-negative real
bf1_age:
  label: Age at first breast feeding
  unit: hours
  scale: positive integer
```

Description of bf data, 8 of 11

```
dc_wt:
  label: Weight at discharge
  unit: kg
  scale: positive real
dc_age:
  label: Age at discharge
  unit: days
  scale: positive integer
```

Description of bf data, 9 of 11

```
dc3_wt:
  label: Weight three days after discharge
  unit: days
  scale: positive real
bf0:
  label: Breastfeeding status at hospital discharge
  values:
    Exclusive: 1
    Partial: 2
    None: 4
```

Description of bf data, 10 of 11

```
bf1:
  label: Breastfeeding status three days after discharge
  values:
    Exclusive: 1
    Partial: 2
    None: 4
bf2:
  label: Breastfeeding status six weeks after discharge
  values:
    Exclusive: 1
    Partial: 2
    None: 4
```

Description of bf data, 11 of 11

```
bf3:
  label: Breastfeeding status three months after discharge
  values:
    Exclusive: 1
    Partial: 2
    None: 4
bf4:
  label: Breastfeeding status six months after discharge
  values:
    Exclusive: 1
    Partial: 2
    None: 4
```

Creating a binary outcome

Bf six months after discharge * breast_feeding_at_six_months Crosstabulation

Count

		breast_feeding_at_six_months		
		.00	1.00	Total
Bf six months after discharge	Exclus.	0	24	24
	Partial	9	0	9
	None	50	0	50
Total		59	24	83

Crosstabulation of predictor and outcome

Binary coding -- control=0, treatment=1 * breast_feeding_at_six_months Crosstabulation

			breast_feeding_at_six_months		
			0	1	Total
Binary coding -- control=0, treatment=1	0=Control	Count	33	12	45
		% within Binary coding -- control=0, treatment=1	73.3%	26.7%	100.0%
	1=Treatment	Count	17	21	38
		% within Binary coding -- control=0, treatment=1	44.7%	55.3%	100.0%
Total		Count	50	33	83
		% within Binary coding -- control=0, treatment=1	60.2%	39.8%	100.0%

Unadjusted odds ratio

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Binary coding -- control=0, treatment=1	1.223	.469	6.795	1	.009	3.397
	Constant	-1.012	.337	9.005	1	.003	.364

a. Variable(s) entered on step 1: Binary coding -- control=0, treatment=1.

Adjusted odds ratio

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Binary coding -- control=0, treatment=1	1.190	.492	5.836	1	.016	3.286
	Mother's age	.008	.037	.047	1	.828	1.008
	Constant	-1.214	.996	1.488	1	.223	.297

a. Variable(s) entered on step 1: Binary coding -- control=0, treatment=1, Mother's age.

Live demo, Risk adjustment

Break #7

- What you have learned
  - Risk adjustment
- What’s coming next
  - Diagnostics

Informal sample size calculations, 1 of 2

- Rule of 50
  - Need 25 to 50 events in each group
  - Based on approximate power calculation
- Example: newborn readmissions for jaundice
  - Occurs about 2% (1/50) of the time
  - Need $25 \times 50 = 1{,}250$ or $50 \times 50 = 2{,}500$ in each group

Informal sample size calculations, 2 of 2

- Rule of 15
  - Need 15 events for each independent variable
  - Smaller ratio implies poor replicability
  - Note: events, not observations

Formal power calculation

Assumptions of logistic regression

- Independence
- Linearity
  - On a log odds scale
- No assumptions about normality

Computing probability estimates, male and 3rd class

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Passenger class			40.639	2	<.001			
	Passenger class(1)	1.319	.212	38.821	1	<.001	3.738	2.469	5.660
	Passenger class(2)	.250	.257	.945	1	.331	1.284	.775	2.127
	Sex(1)	1.528	.199	58.969	1	<.001	4.608	3.120	6.806
	Passenger class * Sex			39.412	2	<.001			
	Passenger class(1) by Sex (1)	1.883	.428	19.323	1	<.001	6.572	2.839	15.214
	Passenger class(2) by Sex (1)	2.229	.417	28.567	1	<.001	9.289	4.102	21.035
	Constant	-2.029	.140	210.940	1	<.001	.132		

a. Variable(s) entered on step 1: Passenger class, Sex, Passenger class * Sex .

- log odds = -2.029
- odds = exp(log odds) = 0.1315
- prob = odds / (1+odds) = 0.1162

Computing probability estimates, male and first class

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Passenger class			40.639	2	<.001			
	Passenger class(1)	1.319	.212	38.821	1	<.001	3.738	2.469	5.660
	Passenger class(2)	.250	.257	.945	1	.331	1.284	.775	2.127
	Sex(1)	1.528	.199	58.969	1	<.001	4.608	3.120	6.806
	Passenger class * Sex			39.412	2	<.001			
	Passenger class(1) by Sex (1)	1.883	.428	19.323	1	<.001	6.572	2.839	15.214
	Passenger class(2) by Sex (1)	2.229	.417	28.567	1	<.001	9.289	4.102	21.035
	Constant	-2.029	.140	210.940	1	<.001	.132		

a. Variable(s) entered on step 1: Passenger class, Sex, Passenger class * Sex .

- log odds = -2.029 + 1.319
- odds = exp(log odds) = 0.4916
- prob = odds / (1+odds) = 0.3296

Computing probability estimates, male and second class

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Passenger class			40.639	2	<.001			
	Passenger class(1)	1.319	.212	38.821	1	<.001	3.738	2.469	5.660
	Passenger class(2)	.250	.257	.945	1	.331	1.284	.775	2.127
	Sex(1)	1.528	.199	58.969	1	<.001	4.608	3.120	6.806
	Passenger class * Sex			39.412	2	<.001			
	Passenger class(1) by Sex (1)	1.883	.428	19.323	1	<.001	6.572	2.839	15.214
	Passenger class(2) by Sex (1)	2.229	.417	28.567	1	<.001	9.289	4.102	21.035
	Constant	-2.029	.140	210.940	1	<.001	.132		

a. Variable(s) entered on step 1: Passenger class, Sex, Passenger class * Sex .

- log odds = -2.029 + 0.25
- odds = exp(log odds) = 0.1688
- prob = odds / (1+odds) = 0.1444

Computing probability estimates, female and third class

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Passenger class			40.639	2	<.001			
	Passenger class(1)	1.319	.212	38.821	1	<.001	3.738	2.469	5.660
	Passenger class(2)	.250	.257	.945	1	.331	1.284	.775	2.127
	Sex(1)	1.528	.199	58.969	1	<.001	4.608	3.120	6.806
	Passenger class * Sex			39.412	2	<.001			
	Passenger class(1) by Sex (1)	1.883	.428	19.323	1	<.001	6.572	2.839	15.214
	Passenger class(2) by Sex (1)	2.229	.417	28.567	1	<.001	9.289	4.102	21.035
	Constant	-2.029	.140	210.940	1	<.001	.132		

a. Variable(s) entered on step 1: Passenger class, Sex, Passenger class * Sex .

- log odds = -2.029 + 1.528
- odds = exp(log odds) = 0.6059
- prob = odds / (1+odds) = 0.3773

Computing probability estimates, female and first class

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Passenger class			40.639	2	<.001			
	Passenger class(1)	1.319	.212	38.821	1	<.001	3.738	2.469	5.660
	Passenger class(2)	.250	.257	.945	1	.331	1.284	.775	2.127
	Sex(1)	1.528	.199	58.969	1	<.001	4.608	3.120	6.806
	Passenger class * Sex			39.412	2	<.001			
	Passenger class(1) by Sex (1)	1.883	.428	19.323	1	<.001	6.572	2.839	15.214
	Passenger class(2) by Sex (1)	2.229	.417	28.567	1	<.001	9.289	4.102	21.035
	Constant	-2.029	.140	210.940	1	<.001	.132		

a. Variable(s) entered on step 1: Passenger class, Sex, Passenger class * Sex .

- log odds = -2.029 + 1.319 + 1.528 + 1.883
- odds = exp(log odds) = 14.8946
- prob = odds / (1+odds) = 0.9371

Computing probability estimates, female and second class

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Passenger class			40.639	2	<.001			
	Passenger class(1)	1.319	.212	38.821	1	<.001	3.738	2.469	5.660
	Passenger class(2)	.250	.257	.945	1	.331	1.284	.775	2.127
	Sex(1)	1.528	.199	58.969	1	<.001	4.608	3.120	6.806
	Passenger class * Sex			39.412	2	<.001			
	Passenger class(1) by Sex (1)	1.883	.428	19.323	1	<.001	6.572	2.839	15.214
	Passenger class(2) by Sex (1)	2.229	.417	28.567	1	<.001	9.289	4.102	21.035
	Constant	-2.029	.140	210.940	1	<.001	.132		

a. Variable(s) entered on step 1: Passenger class, Sex, Passenger class * Sex .

- log odds = -2.029 + 0.25 + 1.528 + 2.229
- odds = exp(log odds) = 7.2283
- prob = odds / (1+odds) = 0.8785

Assessing linearity on a log scale, 1 of 3

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	20.801	8	.008

Contingency Table for Hosmer and Lemeshow Test

		Did the passenger survive? = No		Did the passenger survive? = Yes		
		Observed	Expected	Observed	Expected	Total
Step 1	1	46	52.015	35	28.985	81
	2	41	45.830	33	28.170	74
	3	42	38.768	22	25.232	64
	4	37	41.107	32	27.893	69
	5	52	48.100	30	33.900	82
	6	51	43.422	24	31.578	75
	7	46	41.800	27	31.200	73
	8	43	37.441	23	28.559	66
	9	47	42.683	29	33.317	76
	10	38	51.834	58	44.166	96

Assessing linearity on a log scale, 2 of 3

Assessing linearity on a log scale, 3 of 3

How good are your predictions, 1 of 2

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1155.857[a]	.333	.461

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

How good are your predictions, 2 of 2

Classification Table[a]

			Predicted		
			Did the passenger survive?		
	Observed		No	Yes	Percentage Correct
Step 1	Did the passenger survive?	No	841	22	97.5
		Yes	222	228	50.7
	Overall Percentage				81.4

a. The cut value is .500

Live demo, Diagnostics

Summary

- What you have learned
  - Test of two proportions
  - Chi-square test of independence
  - Odds ratio versus relative risk
  - Concepts behind the logistic regression model
  - Logistic regression with categorical variables
  - Logistic regression with interactions
  - Risk adjustment
  - Diagnostics

Additional topics??

Speaker notes

- Learning objectives
  - Define Logistic Regression analysis
  - Compare and Contrast Multiple Regression analysis and Logistic Regression analysis
  - List and define the three types of logistic Regression analysis
  - Define how predictor variables are selected
  - Define Multicollinearity
  - List and explain the problems that Multicollinearity causes if present in a Logistic Regression analysis
  - List and explain the methods for selecting predictor variables
  - List and explain the problems that too many predictors can cause in a Logistic Regression analysis
  - Compare and contrast Logistic Regression analysis with Linear and Multiple Regression analysis
  - Explain the model equation for a Logistic Regression analysis
  - Explain the problem with the Assumption of Linearity for a Logistic Regression analysis
  - Explain the method that is used to overcome the problem with the Assumption of Linearity
  - Define Maximum Likelihood Estimation
  - Explain Probability
  - Explain Odds
  - Explain the conversion of probability to odds
  - Define and explain the model equation for a Logistic Regression analysis
  - Define reference standard
  - Identify the value that represents the fit of the model
  - Define Log likelihood
  - List and define the 4 different calculations for R squared for an Logistic Regression analysis
  - Define and explain the Wald Statistic
  - Explain the importance of Odds Ratios in a Logistic Regression analysis
  - Identify Exp(B)
  - Identify and define the methods for predictor variable input in a Logistic Regression analysis
  - Understand which predictor variable input method is best used
  - Identify, define and explain the testing for all assumptions
  - Identify and explain all “Important Considerations” for a Logistic Regression analysis
  - Calculate Sensitivity, Specificity, PPV and NPV using the information found in the SPSS Classification Table
  - Explain the meaning of Block 0

- Explain the meaning of Block 0
- Explain the meaning of all subsequent blocks (following Block 0) in a Logistic Regression analysis
- Interpret an Odds Ratio for a continuous predictor variable
- Interpret an Odds Ratio for a categorical predictor variable
- Calculate the sensitivity and specificity of the model
- Explain what an ROC identifies in a Logistic Regression analysis
- Create a complete write up for a Logistic Regression analysis
- Know the table that must be included in a write up for a Logistic Regression analysis