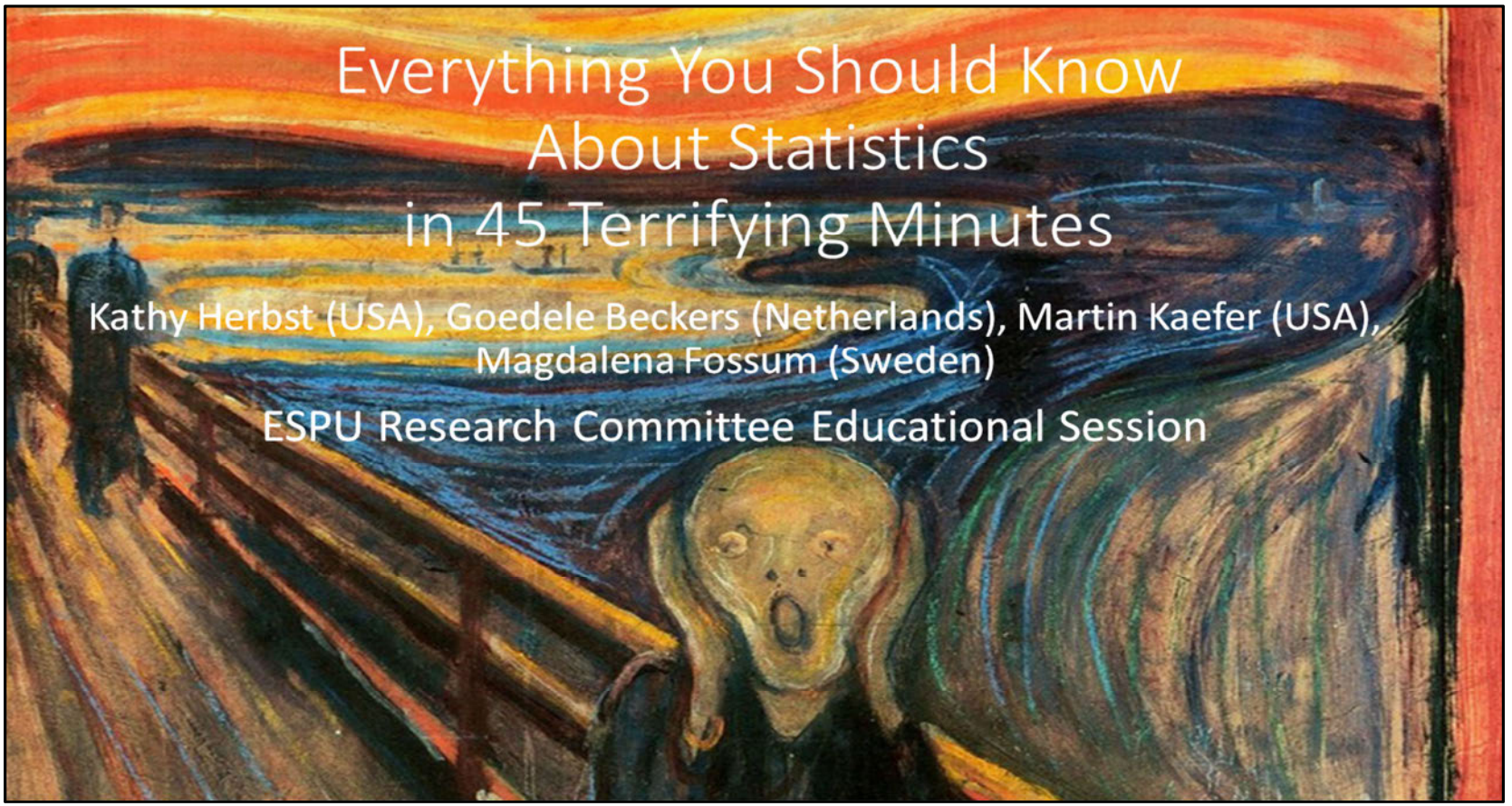
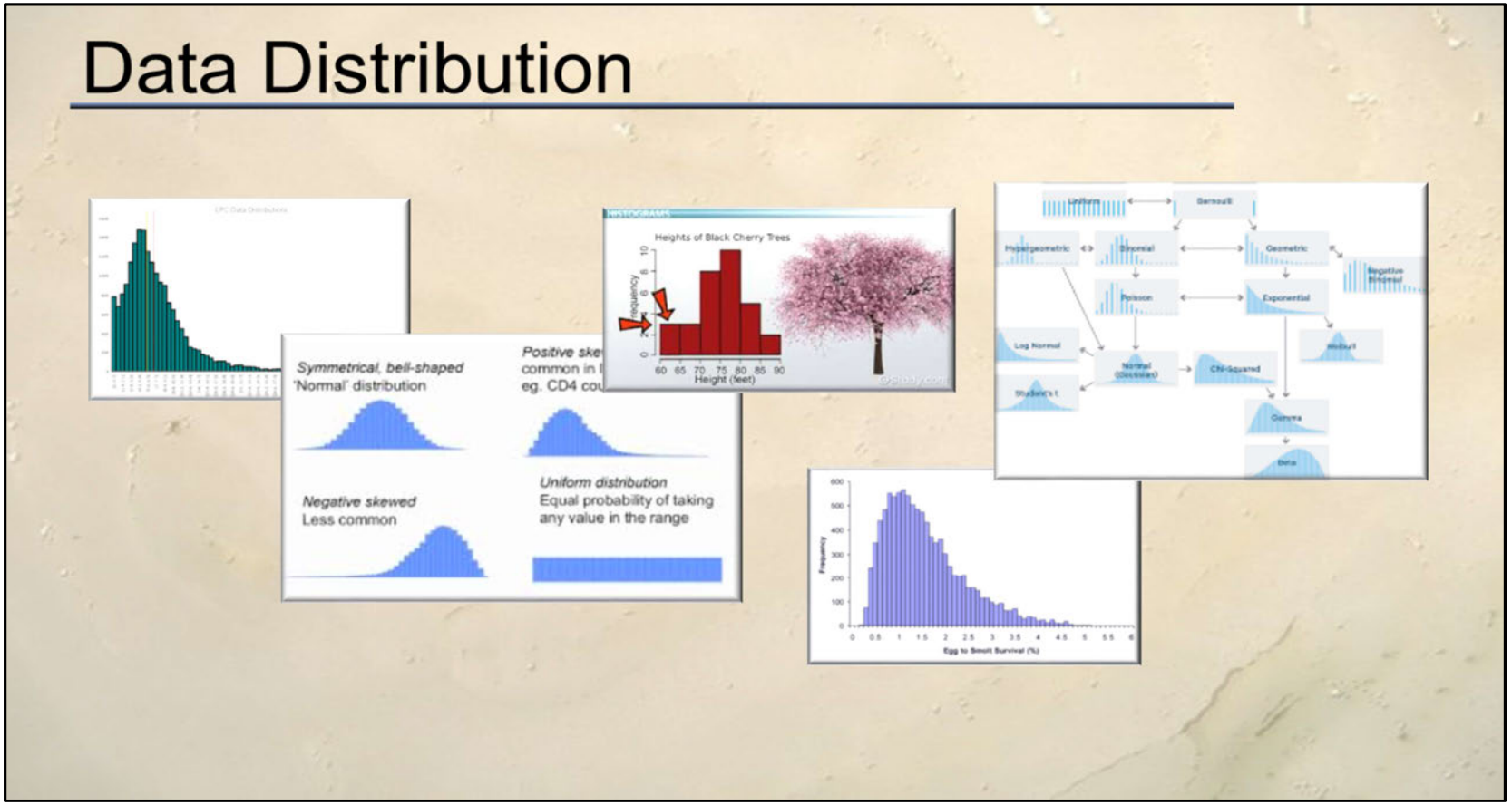


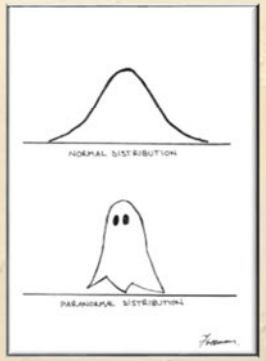
It's impossible to teach everything about statistics in 45 minutes, so our goal is for you to know the some of the basics without having to learn any of the math. This way you'll know what information a statistician will need when working with one, and know some common rules so you can better evaluate the statistics in a manuscript.



I wanted to start with data distribution because that's how I always start. Some of you may think this is rudimentary, but it is very important, and I see distribution described incorrectly all the time in publications. And if you don't understand distribution, there is a good chance you don't understand your data, your analysis will be flawed, or you'll interpret it incorrectly.

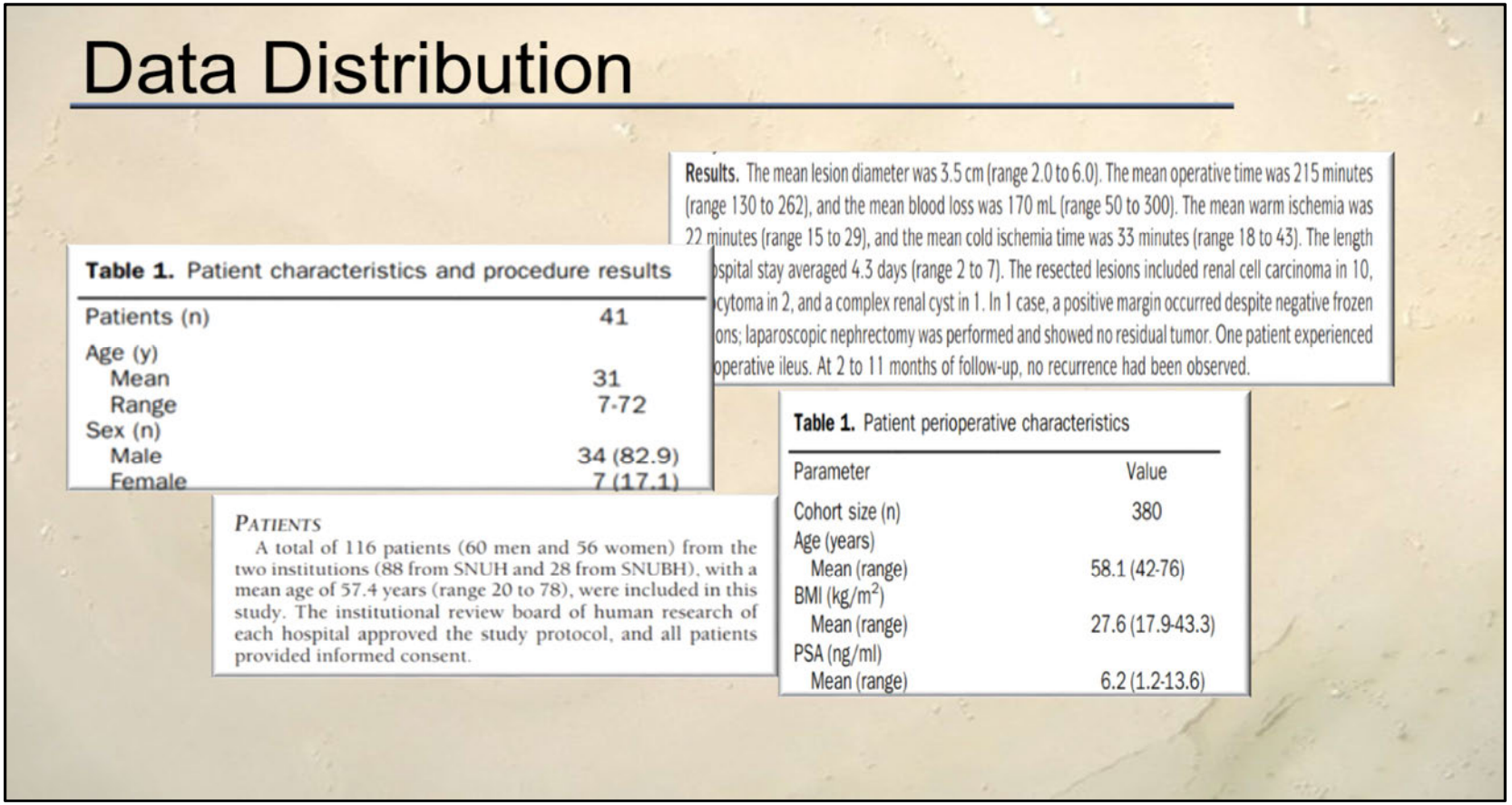
Data Distribution

1. Allows you to understand your cohort
2. Allows you to determine how well the true mean is represented in your data
3. It determines which statistical tests should be used
4. Assists with data interpretation

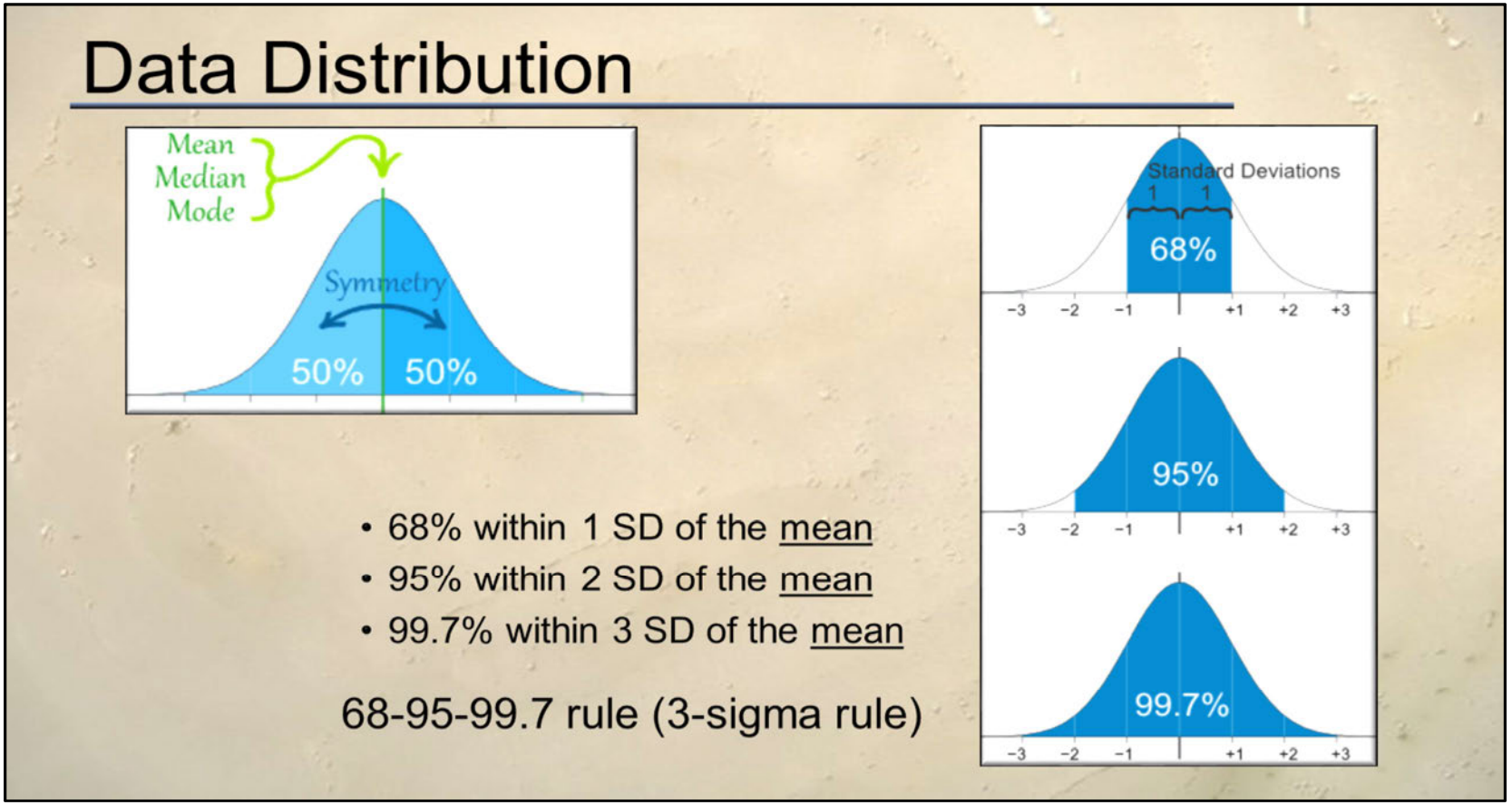


Why is it so important? Knowing the distribution of a variable allows you to understand the sample you're analyzing. For example, if I have data I've collected from a retrospective chart review on kids that underwent reimplantation at my center, and the mean age at time of surgery is 9 years old it tells me one of two things: either I have a built-in selection bias somewhere in my study, probably in my inclusion or exclusion criteria, where the majority of patients I identified were kids who had very delayed presentation or were medically managed for a long time. Or my data is totally flawed as this age doesn't make clinical sense.

Distribution also allows you to determine how your data is dispersed. Is it really spread out so there's a lot of variability between each data point? Is it really narrow? Data dispersion is very important and will be talked about later in this session when we discuss confidence intervals. Distribution determines which statistical tests should be used. There are assumptions about data distribution in most statistical tests. Distribution is important to how the data is interpreted. Does the population under study match your typical patients? Can you apply these findings to your practice?



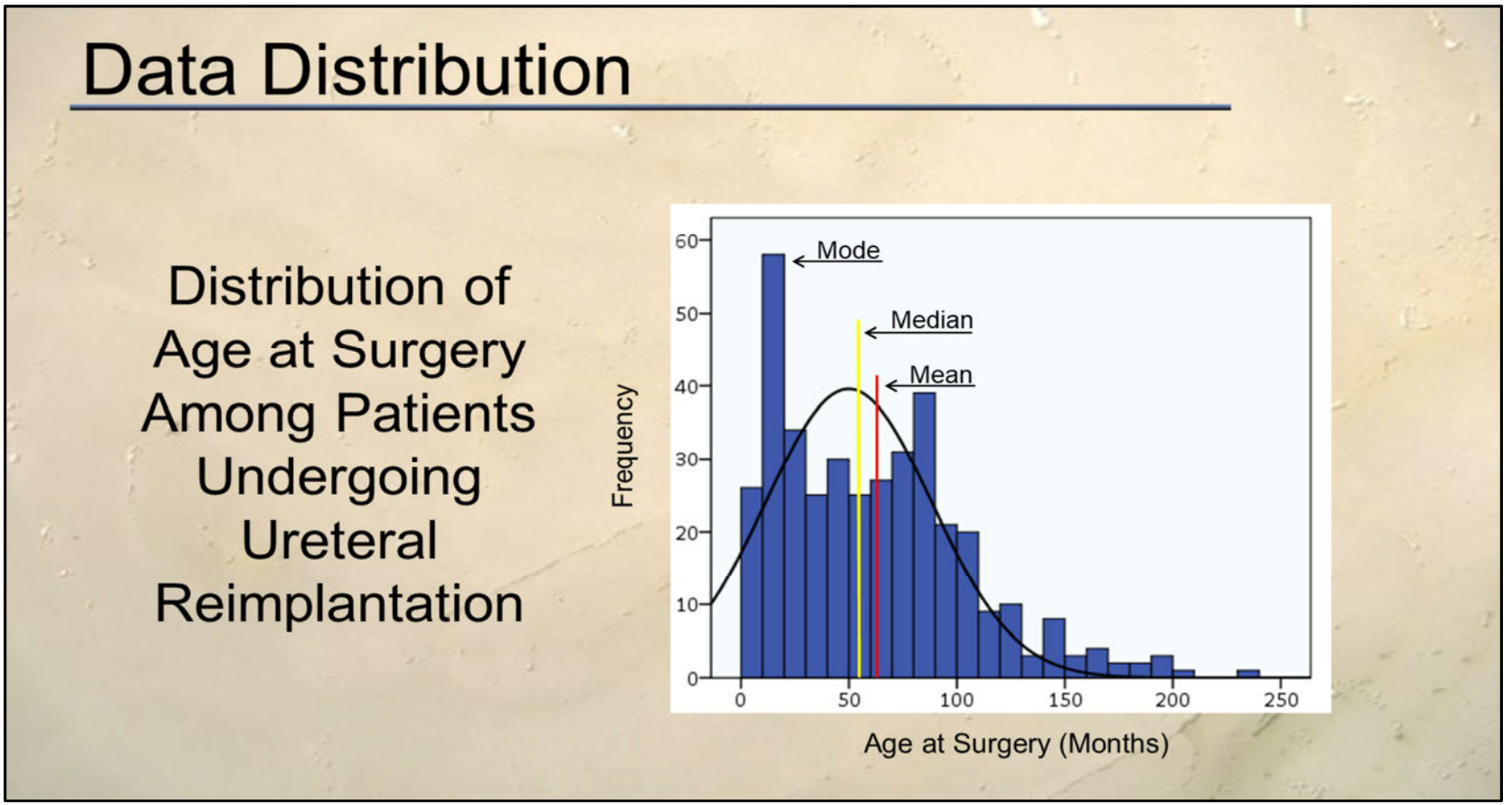
When you use descriptive statistics, you're describing what your data looks like. I often see articles published where authors use the wrong statistics to describe their data. Which makes me think they don't really know what their data looks like or they don't understand statistics. Which, in turn, makes me question their findings. This is also why I felt it important to spend time on distribution.



Let's start with a normal distribution. I'm sure you've all seen this - half of the values are less than the mean, half are more than the mean, and the mean, median, and mode are all the same number.

When describing a normal distribution, you use the mean and the standard deviation. The mean describes the average number for that variable. The standard deviation describes how spread out the data is around the mean: 68% of the values are within 1 standard deviation of the MEAN, 95% within 2 standard deviations, and 99.7% within 3 standard deviations.

So, for a normal distribution, almost all values fall within 3 standard deviations. This is known as the 68-95-99.7 rule, or the 3-sigma rule.



Data are sometimes not normally distributed – they're skewed. Here's a graph of age at time of surgery for a cohort of patients who underwent ureteral reimplantation overlayed with a normal distribution curve. This data is positively skewed or skewed to the right because the long tail is on the positive side of the peak or to the right of the peak.

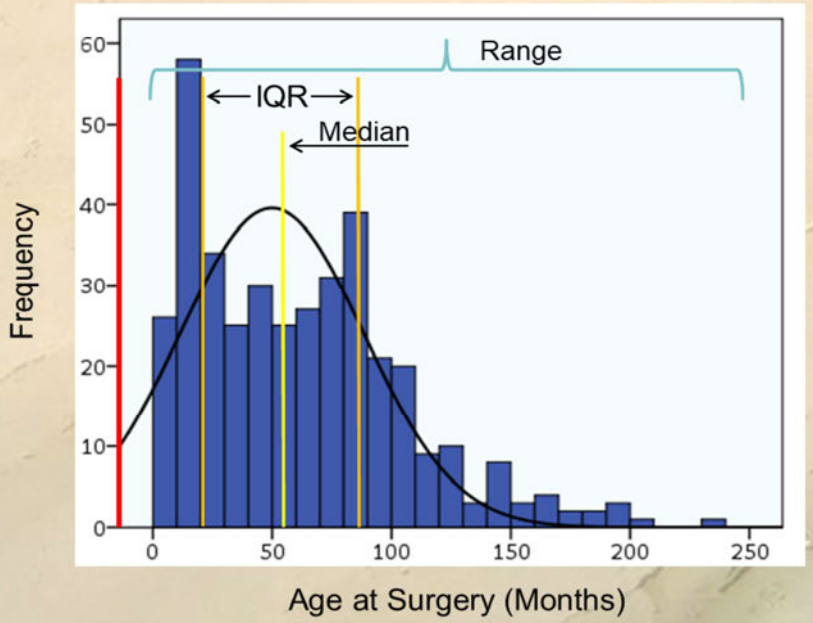
Notice the normal distribution curve goes below 0. The mean is now influenced by extreme values, it's to the right of the peak value, and is no longer a representative measure of the data. If this was a negatively or left skewed distribution, the mean would be to the left of the peak value.

The mode is the peak value – the most frequent value. The median is the middle value. Either the median or the mode are appropriate descriptors. As standard deviation is defined by how far away from the mean data values are, it is not a good measure for skewed data.

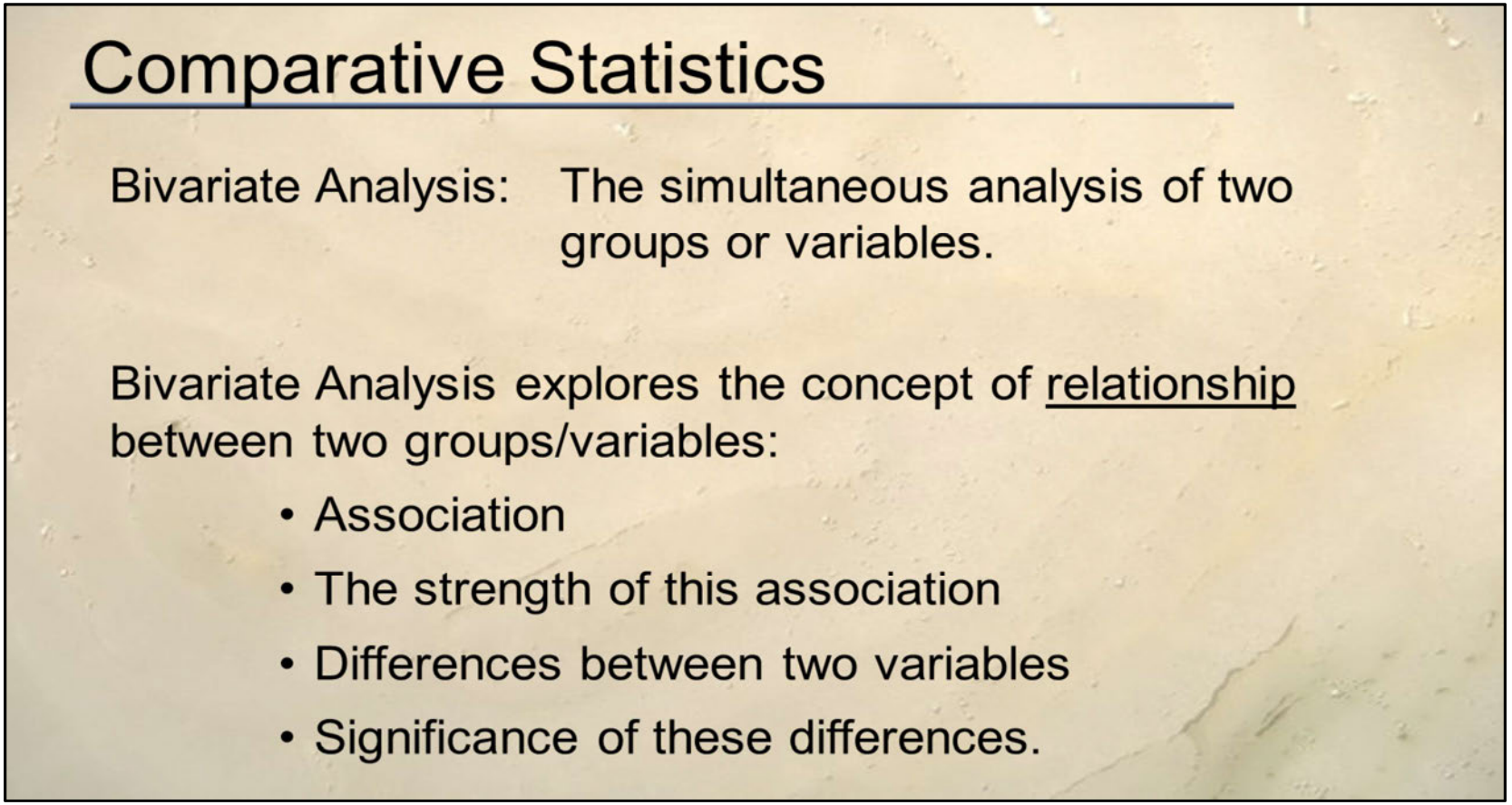
If we were to apply the standard deviation to this data, then the 1st standard deviation to the left would be close to 0, and the 2rd standard deviation to the left would be below 0.

Data Distribution

Distribution of Age at Surgery Among Patients Undergoing Ureteral Reimplantation



Range is a much better measure. The best statistic to describe skewed data dispersion is inner quartile range, or IQR. If you break the data into 4 equal parts, inner quartile range is the 25th percentile to the 75th percentile. It represents 50% of the data. This measure is better than range, as you can tell where the bulk of the data resides. With a range, you don't know where the extreme values are.



Now that we know how distribution works, let's move onto comparative statistic. The next step in most analysis is to conduct bivariate analysis – or analysis comparing two groups. These types of analysis are sometimes referred to as univariate analysis.

Bivariate comparisons explore if there's a relationship between two groups, whether an association exists and the strength of this association, or whether there are differences between two groups, and the significance of these differences. What it won't tell you is if these associations are causal. Often this analysis is used in a Table 1 where patient characteristics are compared between two groups to find differences in them other than the outcome of interest. It's also used to identify confounders or risk factors to be included in multivariate models.

Bivariate Comparison

Patient Characteristics Among Patients Who Underwent VUR Intervention

Characteristic	Total VUR Procedures	Reimplantation	Injection	p-Value
Cohort	14,430	7,045 (49)	7,385 (51)	–
Age at Initial Intervention (yrs)	4.7 (2.5 – 7.2)	4.2 (2.1 – 6.7)	5.2 (2.9 – 7.7)	**0.001**[a]
Female	11,999 (83)	5,605 (80)	6,394 (87)	–
Age at Intervention (yrs)	4.9 (2.8 – 7.3)	4.6 (2.4 – 6.7)	5.3 (3.2 – 7.8	**< 0.001**[a]
Male	2,431 (17)	1,440 (20)	991 (13)	–
Age at Intervention (yrs)	3.2 (1.5 – 6.6)	2.7 (1.3 – 5.6)	4.0 (1.8 – 7.8)	–

Data in table are given as n (%) or median (25^{th}, 75^{th} percentile)
[a] **Mann-Whitney U test**

Herbst K, Corbett ST, Lendvay TS, Caldamone AA. Recent Trends in the Surgical Management of Primary Vesicoureteral Reflux in the Era of Dextranomer/Hyaluronic Acid. J Urol. 2014 May:191(5):1628-1633.

Here's an example from a paper on surgery for VUR. As you can see, there's a significant difference in age at intervention between the group who underwent reimplantation and the group who underwent endoscopic injection.

Bivariate Comparison

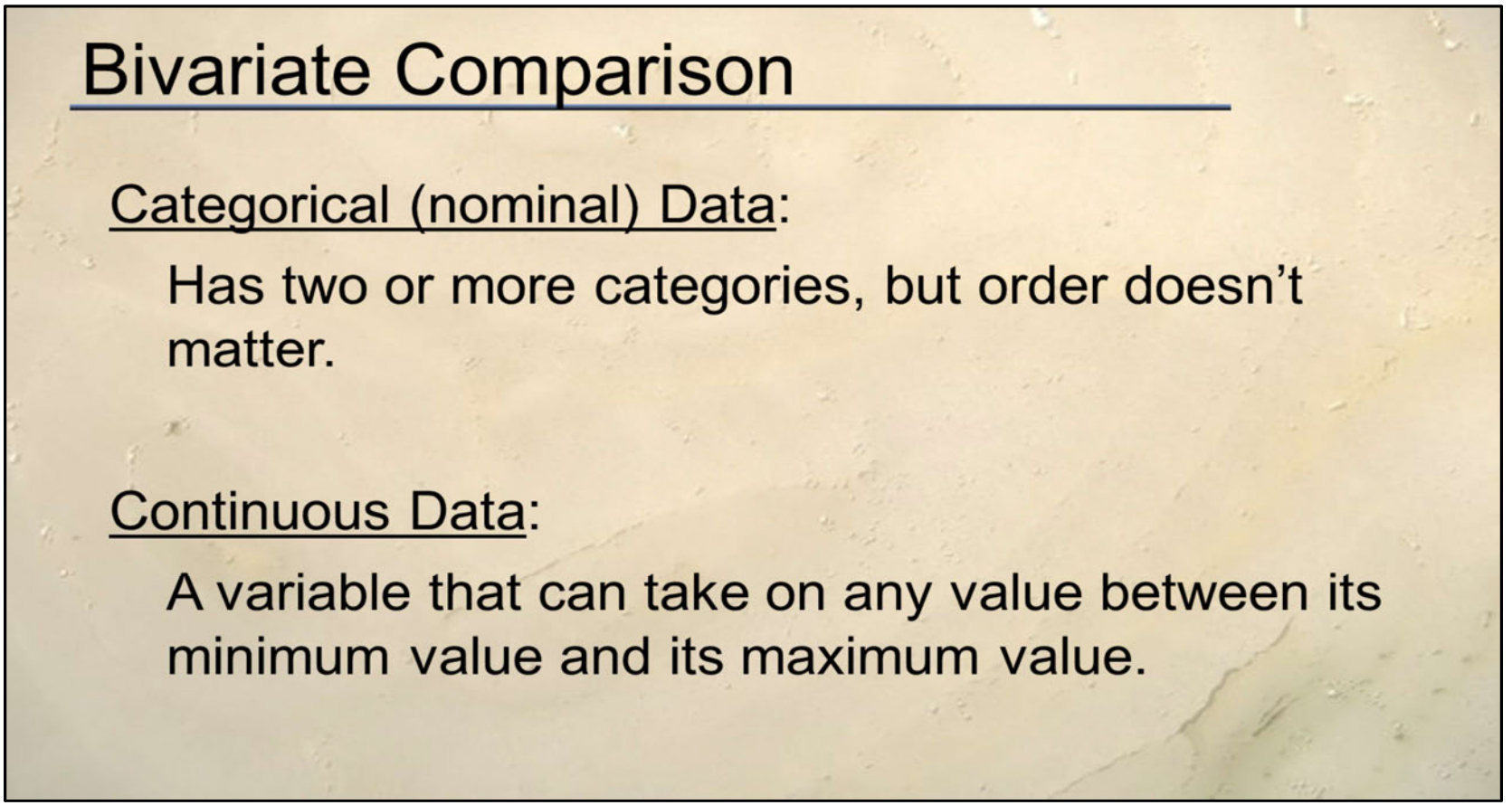
Patient Characteristics Among Patients Who Underwent VUR Intervention

Characteristic	Total VUR Procedures	Reimplantation	Injection	p-Value
Cohort	14,430	7,045 (49)	7,385 (51)	–
Age at Initial Intervention (yrs)	4.7 (2.5 – 7.2)	4.2 (2.1 – 6.7)	5.2 (2.9 – 7.7)	**0.001**[a]
Female	11,999 (83)	5,605 (80)	6,394 (87)	–
Age at Intervention (yrs)	4.9 (2.8 – 7.3)	4.6 (2.4 – 6.7)	5.3 (3.2 – 7.8	< **0.001**[a]
Male	2,431 (17)	1,440 (20)	991 (13)	–
Age at Intervention (yrs)	3.2 (1.5 – 6.6)	2.7 (1.3 – 5.6)	4.0 (1.8 – 7.8)	–

Data in table are given as n (%) or median (25^{th}, 75^{th} percentile)
[a] **Mann-Whitney U test**

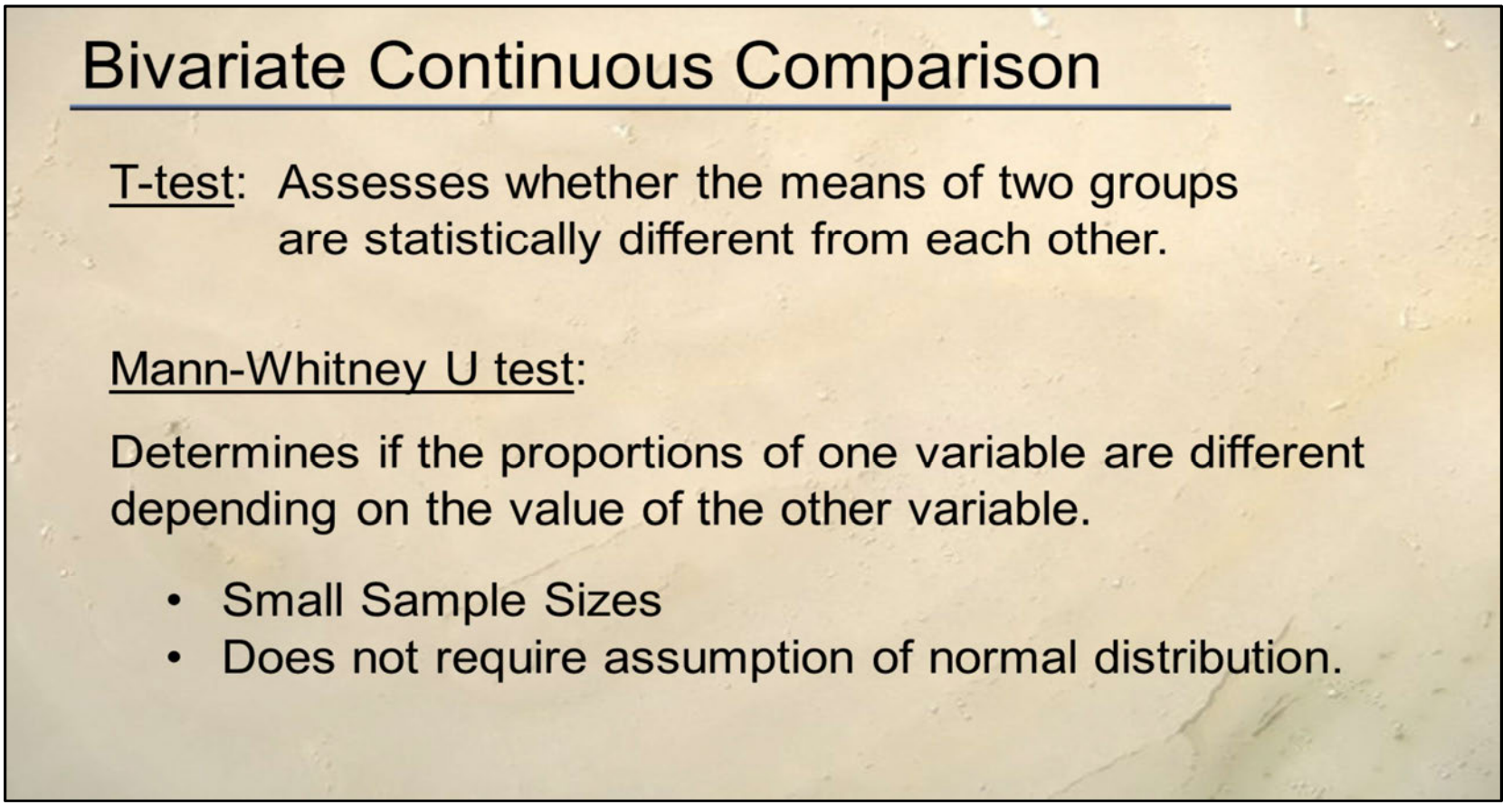
Herbst K, Corbett ST, Lendvay TS, Caldamone AA. Recent Trends in the Surgical Management of Primary Vesicoureteral Reflux in the Era of Dextranomer/Hyaluronic Acid. J Urol. 2014 May:191(5):1628-1633.

But after exploring further by breaking out by gender, you find that this difference is driven by a difference among females and not males. It's important to understand differences in characteristics between the two groups under study as you need to know if it's fair to compare these groups, and these differences may also affect how you interpret your findings.

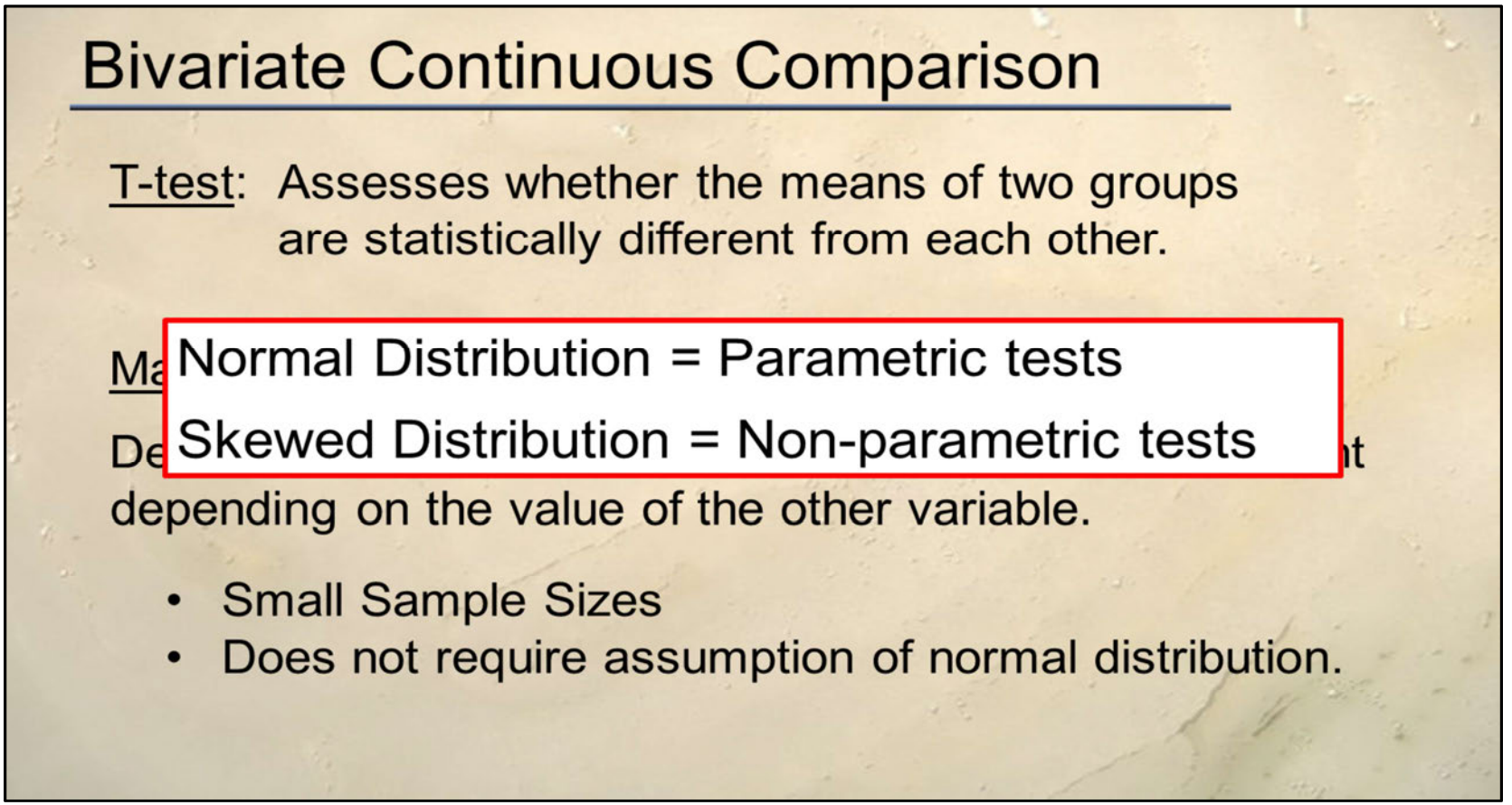


Before determining what statistical test you should use, you need to know what type of variable you're testing. Variables can be classified several ways, but the major classifications are categorical and continuous data.

Categorical data have two or more categories, but order doesn't matter. For example, gender, race, ethnicity. Then there's continuous variables – which is a variable that can take on any value between its minimum value and its maximum value. For example age, weight, or length of follow-up.



Let's talk about continuous variables first. To compare continuous data between two groups, you need to know the distribution of the variable your comparing. If it's normally distributed, then you use a t-test. If your data is skewed, use the Mann-Whitney U test, which is the non-parametric equivalent of a t-test. Parametric statistical tests assume a normal distribution, and non-parametric tests don't assume anything about the distribution so are applicable to skewed data.



Pay attention to how I just said that. Data is not parametric or non-parametric – statistical tests are. Data is normally distributed or skewed.

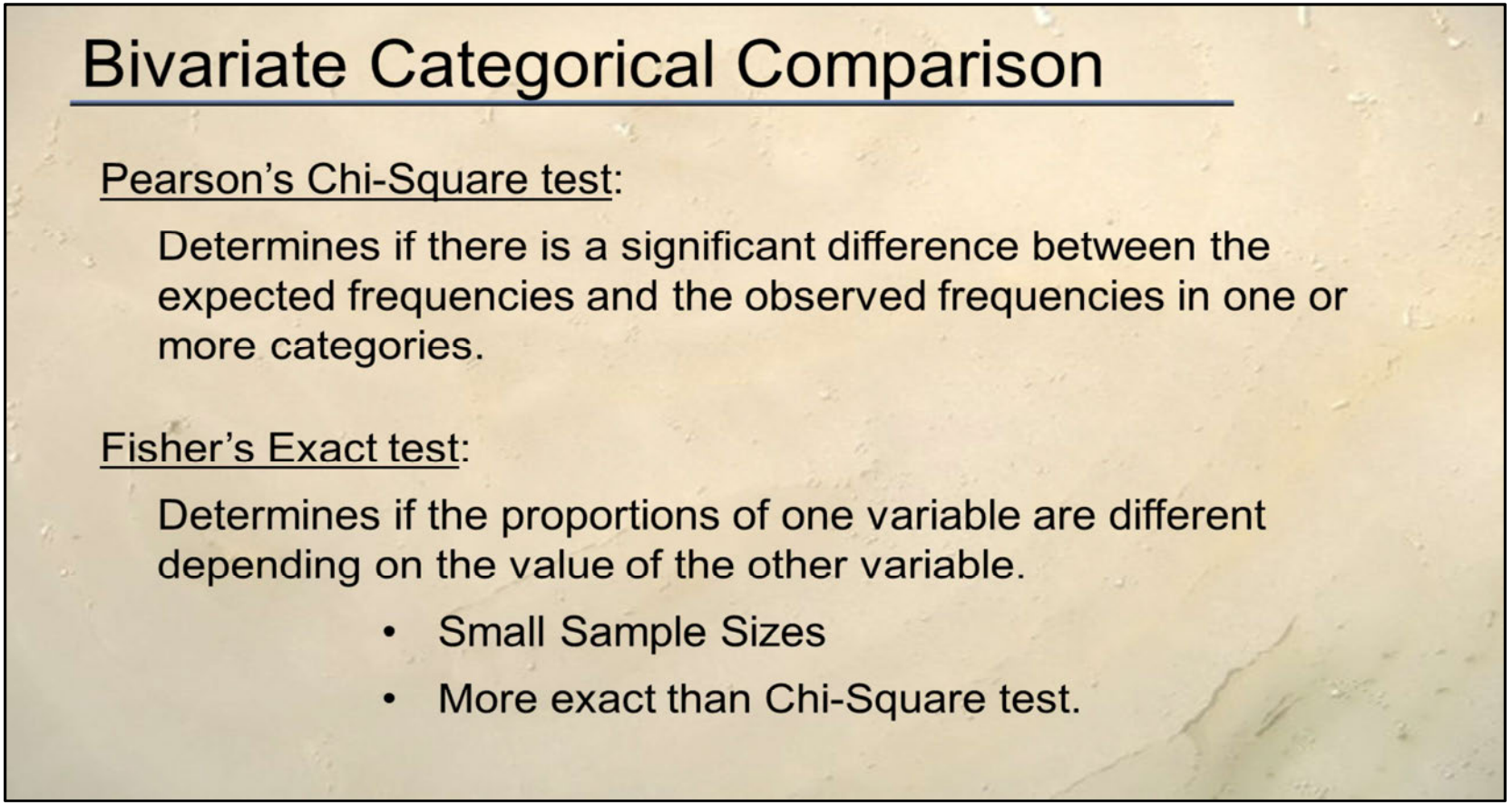
Bivariate Categorical Comparison

	Characteristic	
Exposure	Yes	No
Yes		
No		

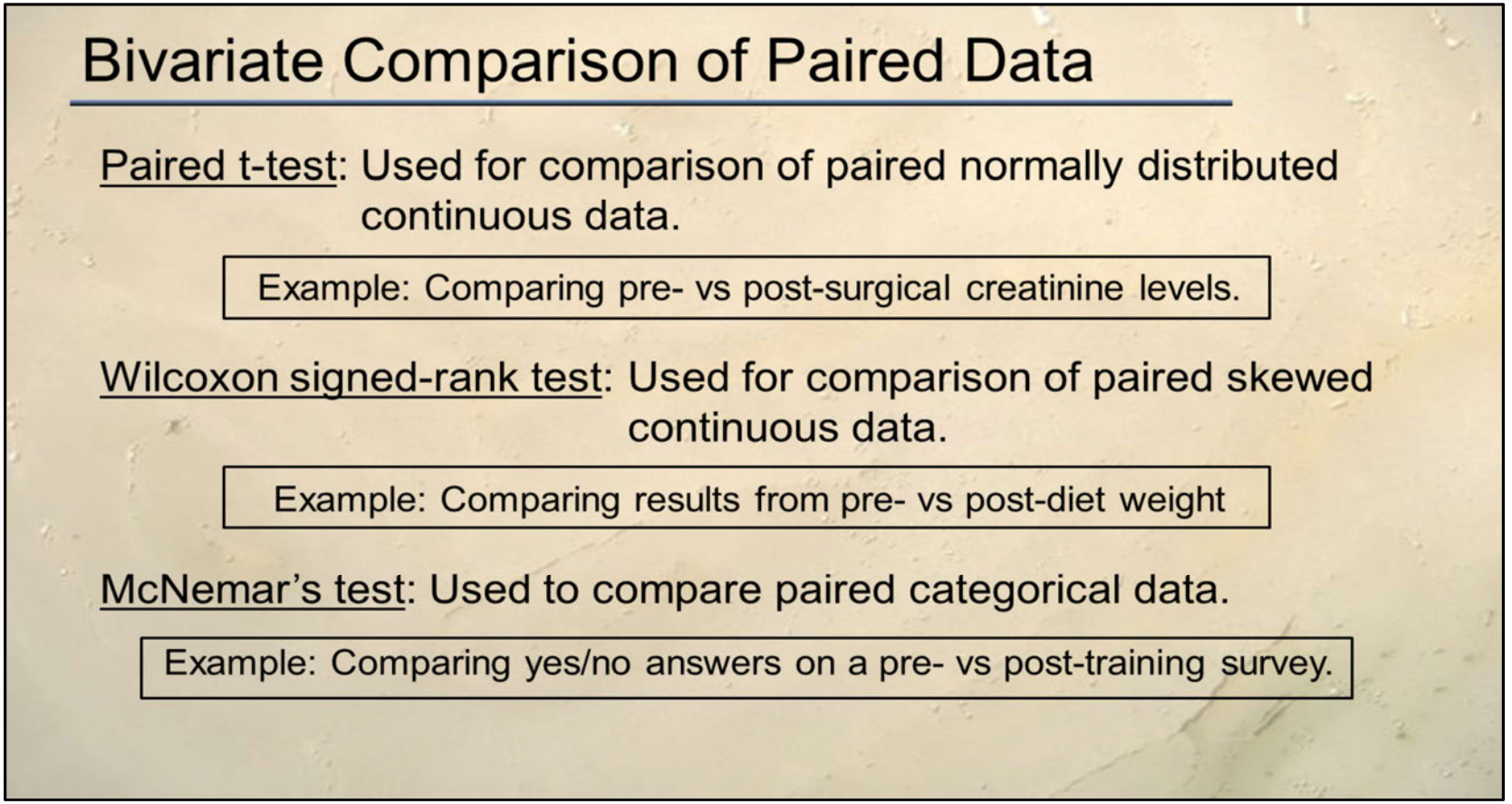
Do females ≤18 years experience more UTIs than males?

	Gender (Female)	
UTI	Yes	No
Yes	95	16
No	87	25

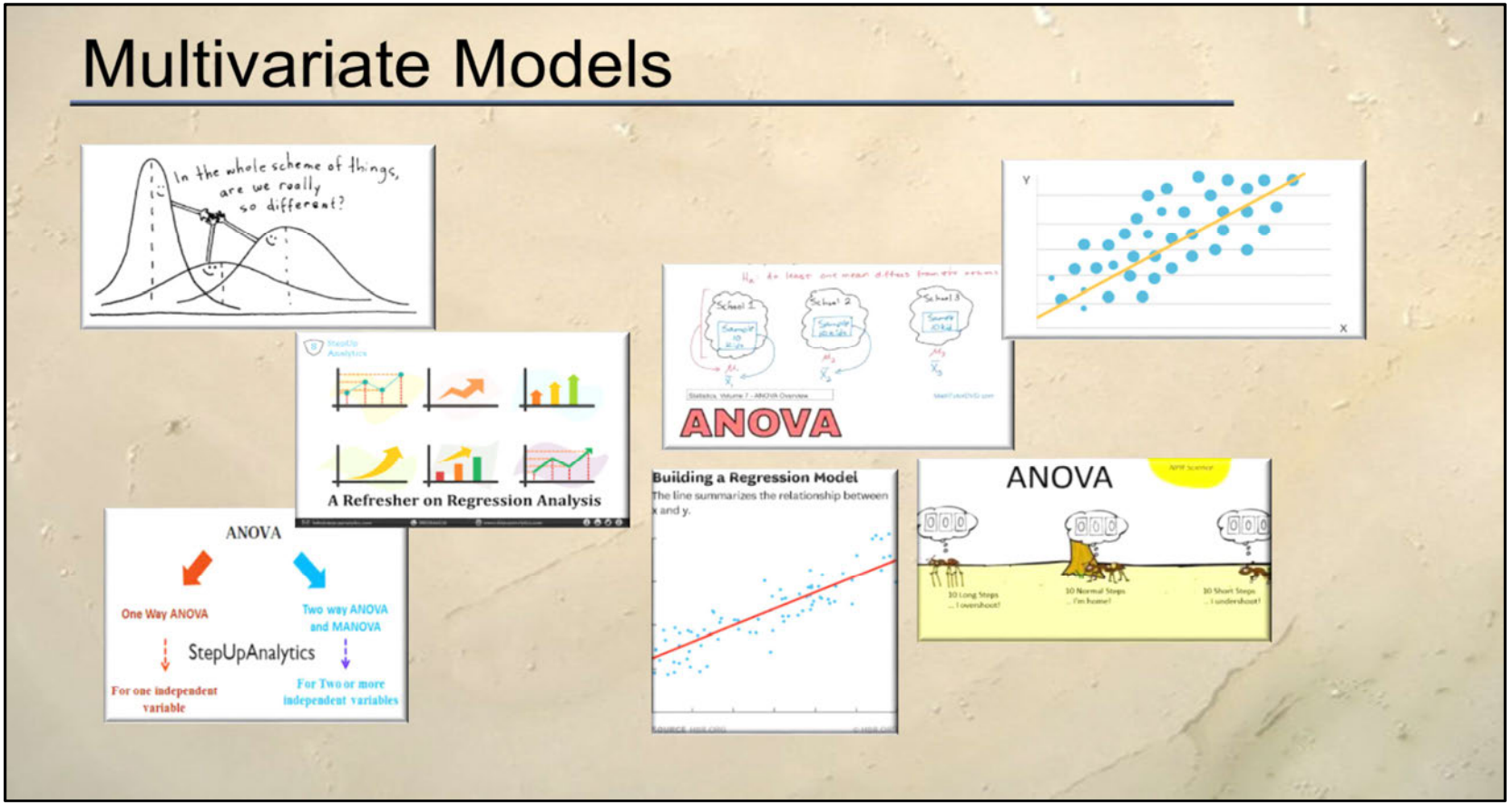
Okay, onto categorical comparisons. I always draw tables when thinking about categorical analysis – they help me clarify my hypothesis and the comparisons I want to make.



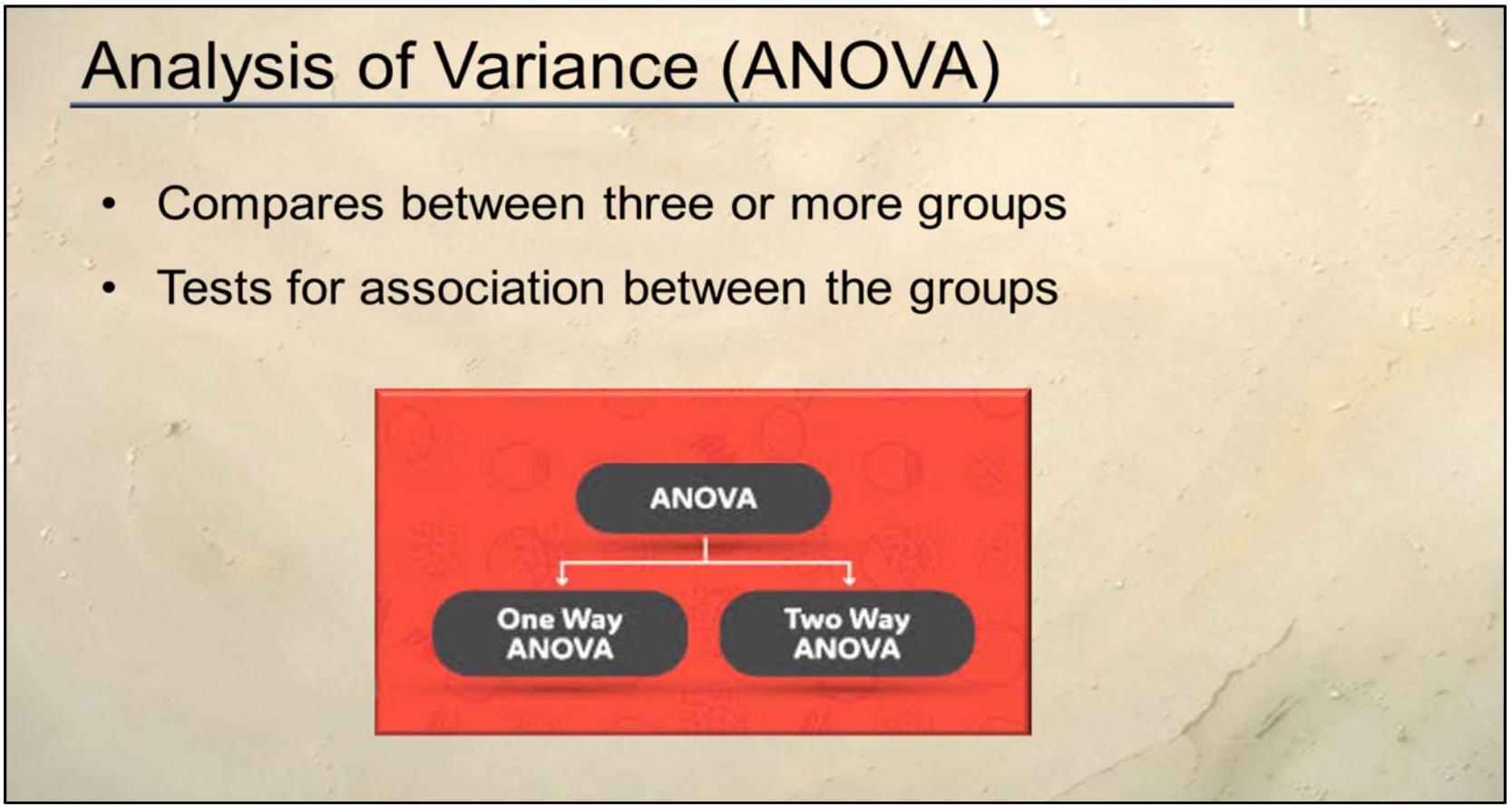
The most common statistical tests for categorical data are Chi-square and Fisher's Exact test.
The chi-squared test uses variance in its formula, so should be used for larger sample sizes. Fisher's Exact test is used if you have a small sample size or a small number of events. By events, I mean an outcome. For example, if my outcome is mortality, and 3 subjects die in my study, then my event count is 3. If you're unsure what to use, use the Fisher's Exact test – you can't go wrong.



Finally, be careful your data isn't paired data. For example, if I'm comparing findings from renal ultrasounds that patients had before surgery to those that they had after surgery, the ultrasounds are paired because they happened to the same patient. Bivariate tests for paired data are different than those for unpaired data.



What if your data has three or more categories or outcomes? For example, kids with VUR could get worse, remain stable, or resolve after surgery. Or what if want to see if a characteristic changes your outcome? Or you may want to predict how likely it is for an outcome to occur given a characteristic. This is where multivariate models come in handy. There are several models that can be used, the most common are analysis of variance or ANOVA and regression.



ANOVA is used to compare 3 or more groups. For example, if I wanted to compare outcomes for kids who underwent reimplantation vs endoscopic injection, and my outcome was that their VUR got worse, stabilized, got better, or totally resolved, I would use ANOVA.

But what's the difference between one-way ANOVA and two-way ANOVA? To answer that, you need to know what an independent variable is and what a dependent variable is.

Independent vs Dependent Variables

Independent Variable: Not influenced by anything

Dependent Variable: Depends on something

"(Independent Variable) causes a change in (Dependent Variable) and it isn't possible that (Dependent Variable) could cause a change in (Independent Variable)"

An independent variable can stand on its own – it's not influenced by anything.

A dependent variable is just like it sounds – it depends on something.

If you're having trouble figuring out which is the independent variable and which is the dependent variable, try putting them into a sentence, "(Independent variable) causes a change in (Dependent Variable) and it isn't possible that (Dependent Variable) could cause a change in (Independent Variable)." For example, drinking beer causes me to talk loudly but it isn't possible that talking loudly causes me to drink beer. Independent variables are often call predictor variables.

Analysis of Variance (ANOVA)

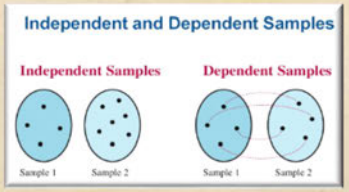
One-way ANOVA: 1 dependent variable and 1 independent variable.

Two-way ANOVA: 1 dependent variable and >1 independent variable.

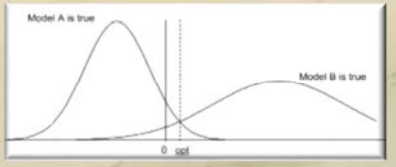
Assumptions

- A normal distribution.

- The samples must be independent from each other.



- The variances in the groups must be equal.



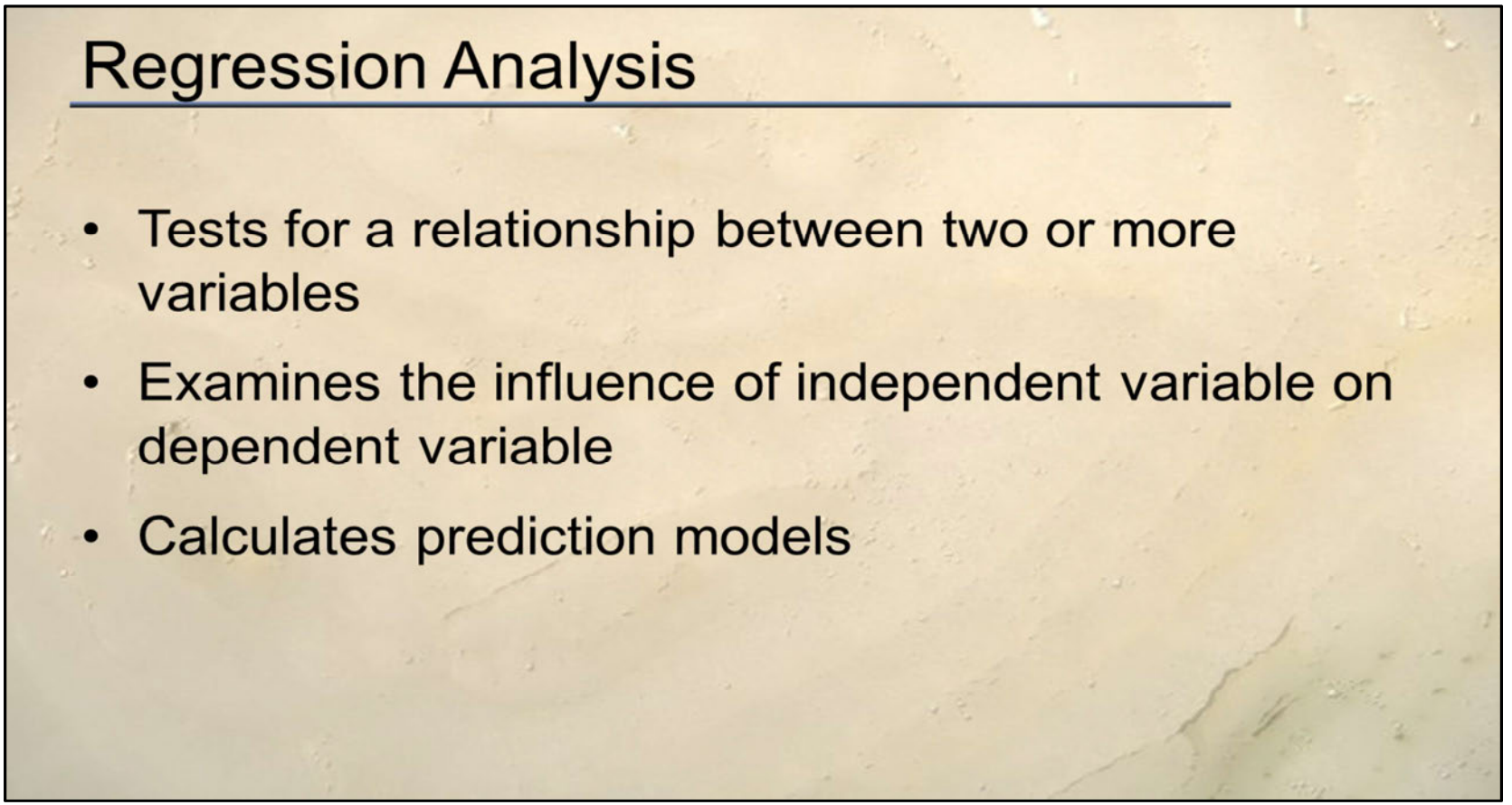
So, now that you know that, the answer to what's a one-way ANOVA and what's a two-way ANOVA is easy. A one-way ANOVA has one independent variable and a two-way ANOVA has two or more independent variables.

There's some assumptions for ANOVA that you need to keep in mind: that the variables are normally distributed, that the samples are independent, and that the variances are equal.

You should be able to tell if the variables are normally distributed now that you know what to look for. What does it mean that the samples are independent? It means that the values of one sample are not dependent on the values of the other sample.

So, if you take a random sample of people and split it into two groups, it's highly likely that the samples are independent. However, if you take a random sample of people, then pick other people to compare them to by matching on age and gender, then the samples are dependent – the matching sample depends on the first sample.

Finally, what does equal variances mean? We discussed variance when we talked about distribution. It means that one sample can't be really spread out and one sample narrowly distributed.



Next, let's talk about regression. Regression is used to explore the relationship between two or more variables, to examine the influence of two or more independent variables on a dependent variable, or to predict something. There's several types of regression analysis, but we're only going to talk about two: linear regression and logistic regression.

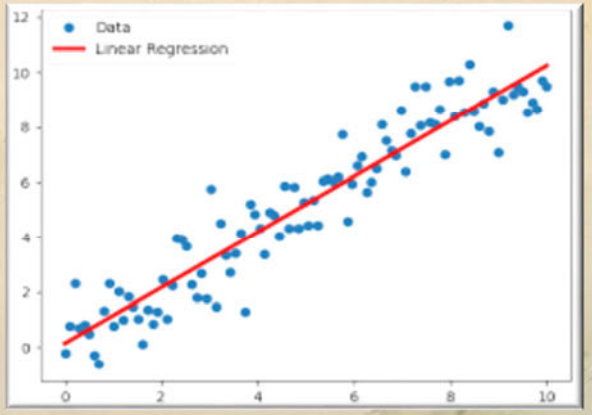
Linear Regression

Simple Linear Regression: 1 dependent and 1 independent

Multiple Linear Regression: 1 dependent and >1 independent

Assumptions

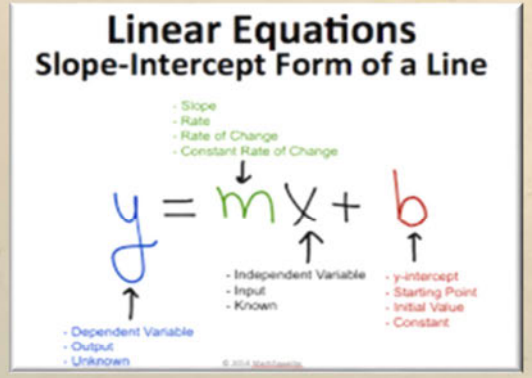
- Variables compared MUST have a linear relationship
- Both variables MUST be continuous

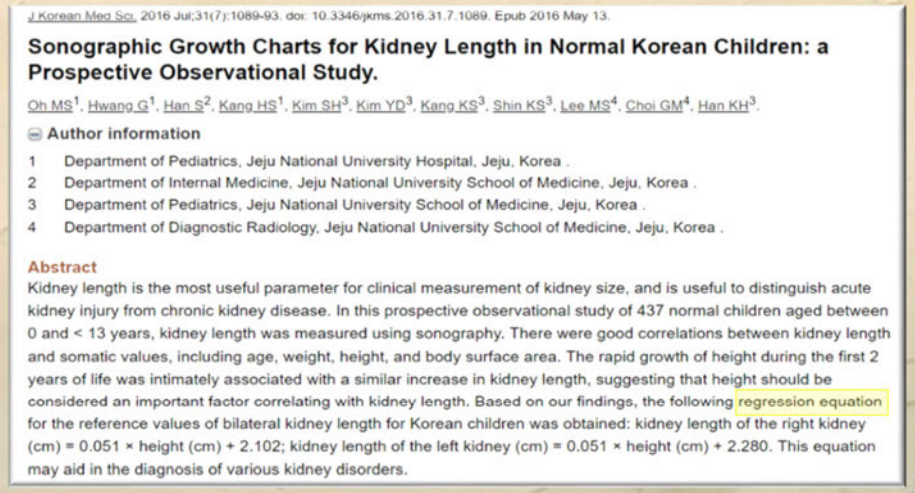


You can only use linear regression if there's a linear relationship between the independent and dependent variable, and both are continuous variables. If you only have one independent variable and one dependent variable, the analysis is called "simple linear regression". If you have multiple independent variables and one dependent variable, the analysis is called multivariate linear regression or multiple linear regression.

Multiple Linear Regression

- Adjusts for confounding effects
- Create prediction models





J Korean Med Sci. 2016 Jul;31(7):1089-93. doi: 10.3346/jkms.2016.31.7.1089. Epub 2016 May 13.

Sonographic Growth Charts for Kidney Length in Normal Korean Children: a Prospective Observational Study.

Oh MS[1], Hwang G[1], Han S[2], Kang HS[1], Kim SH[3], Kim YD[3], Kang KS[3], Shin KS[3], Lee MS[4], Choi GM[4], Han KH[3].

Author information

1 Department of Pediatrics, Jeju National University Hospital, Jeju, Korea .
2 Department of Internal Medicine, Jeju National University School of Medicine, Jeju, Korea .
3 Department of Pediatrics, Jeju National University School of Medicine, Jeju, Korea .
4 Department of Diagnostic Radiology, Jeju National University School of Medicine, Jeju, Korea .

Abstract

Kidney length is the most useful parameter for clinical measurement of kidney size, and is useful to distinguish acute kidney injury from chronic kidney disease. In this prospective observational study of 437 normal children aged between 0 and < 13 years, kidney length was measured using sonography. There were good correlations between kidney length and somatic values, including age, weight, height, and body surface area. The rapid growth of height during the first 2 years of life was intimately associated with a similar increase in kidney length, suggesting that height should be considered an important factor correlating with kidney length. Based on our findings, the following regression equation for the reference values of bilateral kidney length for Korean children was obtained: kidney length of the right kidney (cm) = 0.051 × height (cm) + 2.102; kidney length of the left kidney (cm) = 0.051 × height (cm) + 2.280. This equation may aid in the diagnosis of various kidney disorders.

Multiple regression lets you understand the effect of an independent variable on the dependent variable accounting for the effects of other variables – so it adjusts for confounding.

Another reason to use linear regression is it can create prediction models to estimate outcomes. For example, if I want to estimate kidney size based on age, I can run a linear regression with data where I know the age of the subject and their actual kidney size to calculate a slope-intercept equation. From this equation, I can calculate estimated kidney size based on age. I may even want to add height to my model to adjust for it, especially if it's for the pediatric population as kids with similar ages grow at different rates.

Accuracy of Prediction Models

Teaching Cohort: Sample where you go the prediction model from.

Validation Cohort: Sample population where you test the model.

J Med Imaging (Bellingham). 2018 Apr;5(2):021219. doi: 10.1117/1.JMI.5.2.021219. Epub 2018 Mar 1.

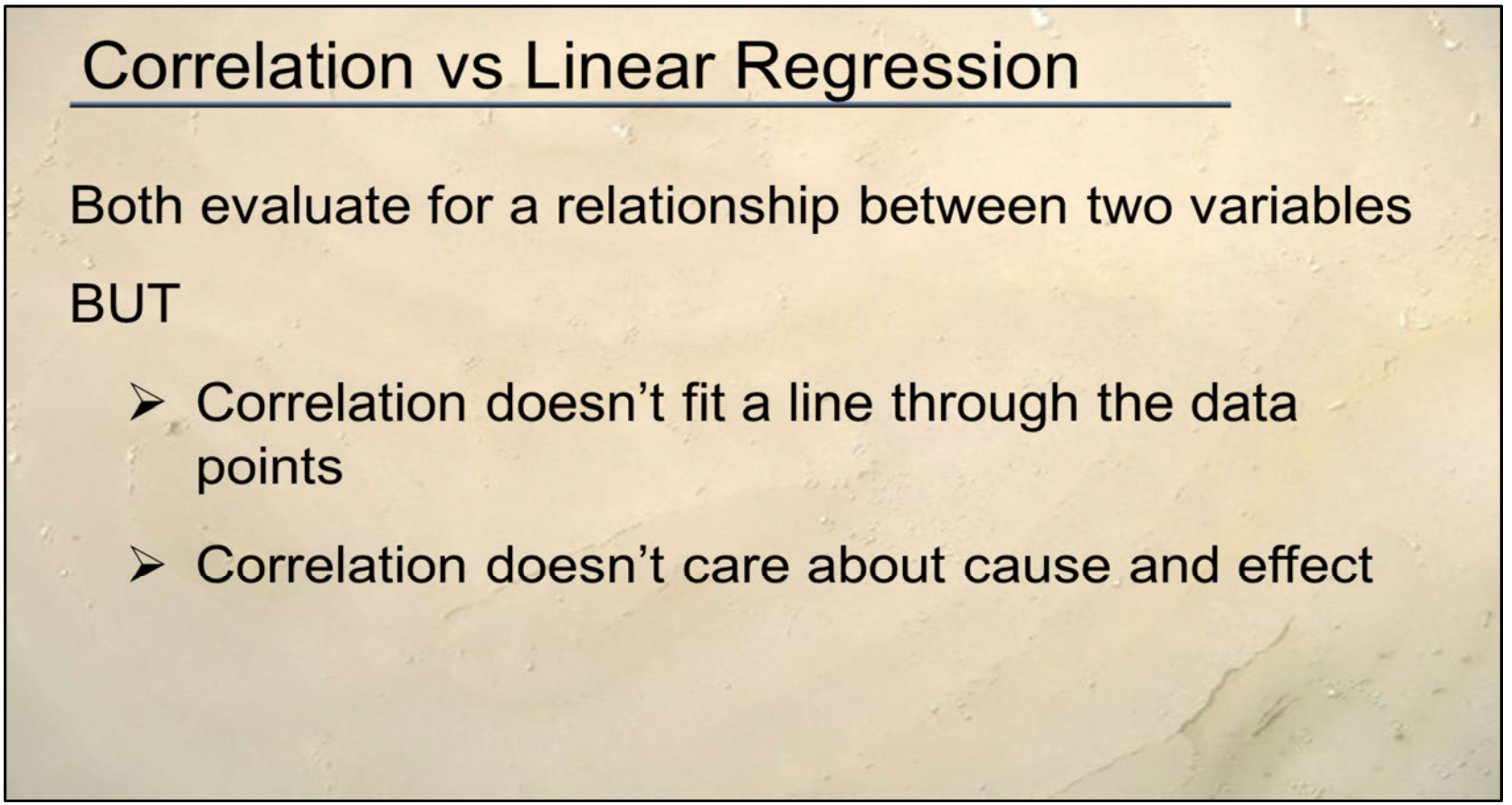
Radiomic signature of infiltration in peritumoral edema predicts subsequent recurrence in glioblastoma: implications for personalized radiotherapy planning.

Rathore S[1,2], Akbari H[1,2], Doshi J[1,2], Shukla G[1,3], Rozycki M[1,2], Bilello M[1,2], Lustig R[4], Davatzikos C[1,2].

Table 2 Performance of the proposed infiltration prediction model for discovery and replication cohorts.

	Discovery cohort	Replication cohort
Odds ratio	10.22	13.66
Accuracy	87.51	89.54
Sensitivity	80.65	97.06
Specificity	87.63	76.73
Balanced accuracy	85.00	87.00
AUC	0.83	0.91

One thing I want to mention about prediction modeling is the use of teaching cohorts and validation cohorts. A teaching cohort, or a discovery cohort, is the sample of people that you derived the prediction model from. The model is always more accurate when applied to these people, because that's where you got the equation. Whenever you're deriving a prediction model or equation, it's a good idea to test the model in a different group of people, a validation or replication cohort, to make sure it holds up. This includes validating findings from ROC curves which are often used to set thresholds for diagnostic testing.



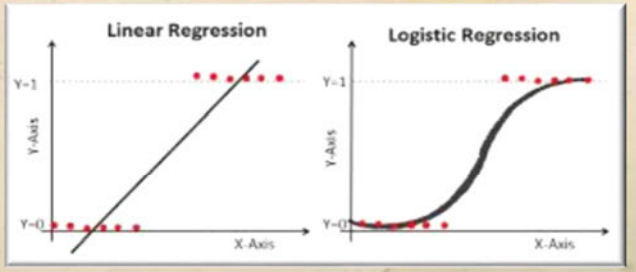
Correlation is another way to evaluate the relationship between two variables. It can help you determine the strength of a relationship, but it doesn't fit a line through the data points. It also doesn't matter which variable is on the X axis and which is on the Y axis, so you don't have to think about cause and effect, or which variable is dependent and which is independent. It's good for exploratory analysis, but running a linear regression is more robust.

Logistic Regression

- Doesn't require a linear relationship
- Doesn't require equal variances
- Often require larger sample sizes

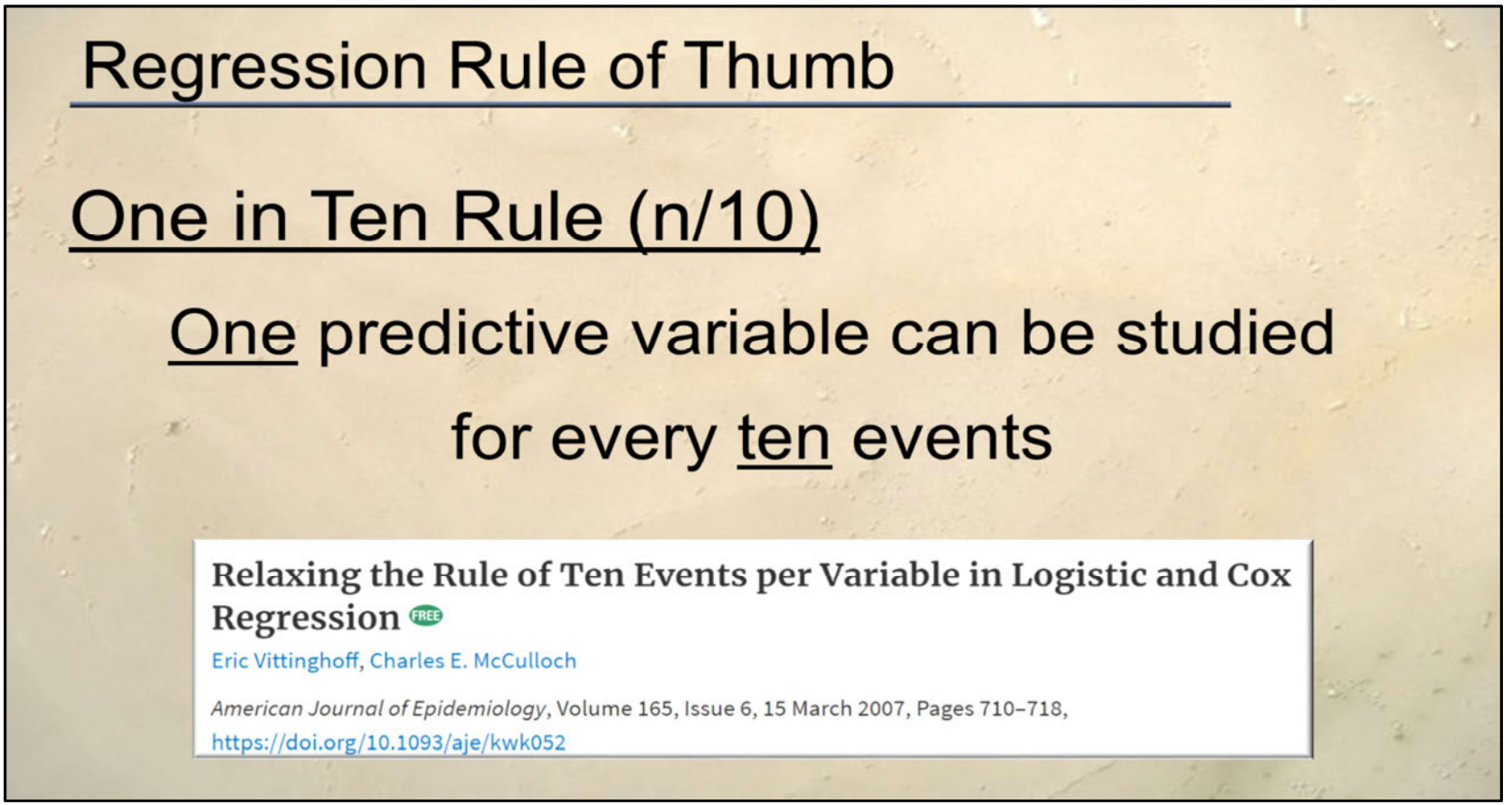
Assumptions

- Dependent variable is binary
- Data isn't matched or paired
- Independent variables are not highly correlated



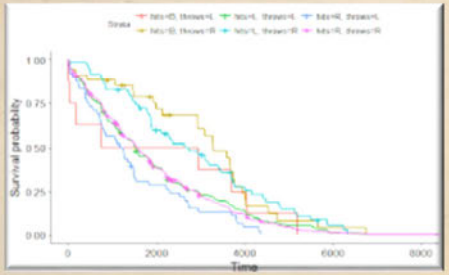
Logistic regression can be used when there's not a linear relationship between two variables or if variances aren't equal. But the dependent variable has to be binary – which means yes or no.

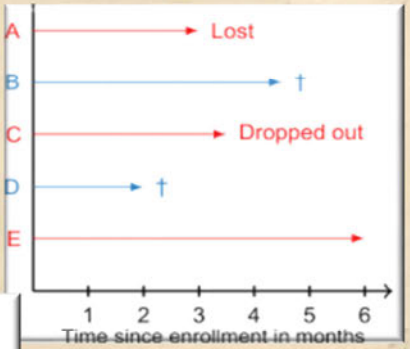
For example, mortality. You're either dead or you're not. Or pregnancy. You're either pregnant or you're not. It also requires that the data isn't matched or paired data, that the independent variables aren't highly correlated with each other - which is a sign of a linear relationship, and it usually requires larger sample sizes. Just like linear regression, logistic regression is good for identifying risk factors. For example, does body weight, calorie intake, fat intake, and age have an influence on the probability of having a heart attack?



There's a rule of thumb you need to know when conducting any type of regression, including Cox regression. It's called the one-in-ten rule or the n/10 rule. This rule states that you should only have one independent or predictor variable in the model for every ten events. So, if I'm looking at survival, and 20 people died in my cohort, then I should only include 2 predictor variables in my regression model. Some studies say that this rule is too conservative. However I look for it, and if it's broken, I look for model fit statistics to make sure the model isn't "overfitted" which means it has too many predictive variables.

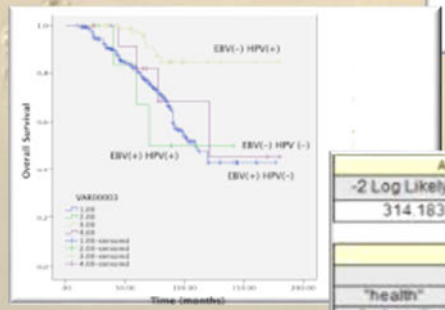
Analysis that Adjust for Time

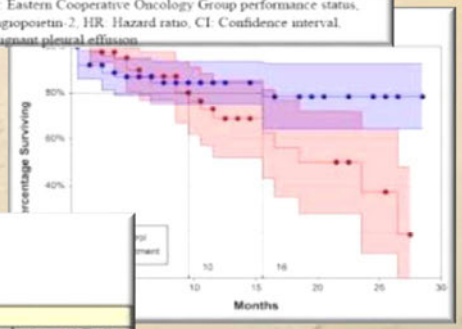




Characteristic	*P*	HR	95% CI for HR	
			Lower	Upper
Gender	0.86	1.06	0.57	1.96
Age	0.21	1.38	0.84	2.26
Stage	0.81	0.92	0.49	1.74
Smoking	0.78	0.92	0.51	1.66
ECOG PS	0.21	0.72	0.32	1.08
MPE control	0.03	0.42	0.19	0.89
Pleural effusion Ang-2	0.01	1.15	1.01	1.32
Serum Ang-2	0.08	1.02	0.99	1.04

ECOG PS: Eastern Cooperative Oncology Group performance status, Ang-2: Angiopoietin-2, HR: Hazard ratio, CI: Confidence interval, MPE: Malignant pleural effusion





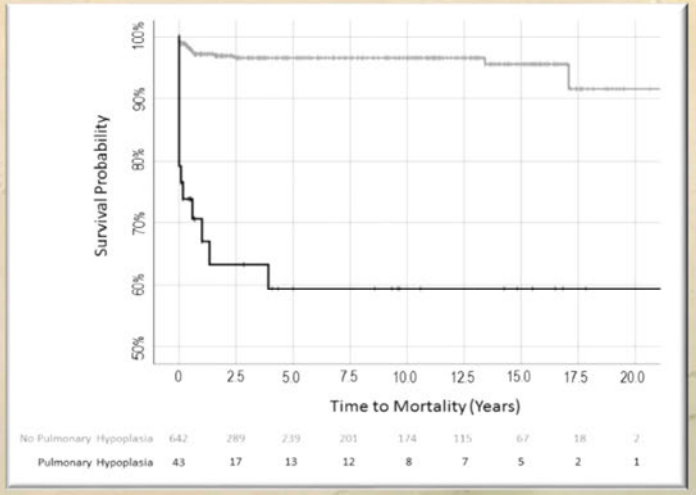
Analysis of -2 Log Likelihood Estimates			
-2 Log Likelyhood	Chi-Square	DF	Pr > ChiSq
314.1839	30.5707	5	0

Analysis of parameter estimates						
	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
"health"	1	-0.0321	0.0076	17.9497	0	0.9685
"volume"	1	-0.0992	0.1938	0.2619	0.6088	0.9056
"pattern"	1	1.0536	0.4379	5.789	0.0161	2.8679
"Duke"	1	-0.1979	0.2503	0.6254	0.4291	0.8204
"Nm23"	1	-0.2877	0.1581	3.3131	0.0687	0.75

Kaplan Meier Estimator

- Non-parametric
- Estimated probability of “event free” survival

	No Hypoplasia		Hypoplasia	
Time Period	Prob	SE	Prob	SE
2.5 years	0.965	0.003	0.633	0.081
5 years	0.965	0.009	0.594	0.085
7.5 years	0.965	0.009	0.594	0.085
10 years	0.965	0.009	590.4	0.085



95% CI for Survival

0.633 + 0.081 = 0.713

0.633 – 0.081 = 0.552

Probability of Death

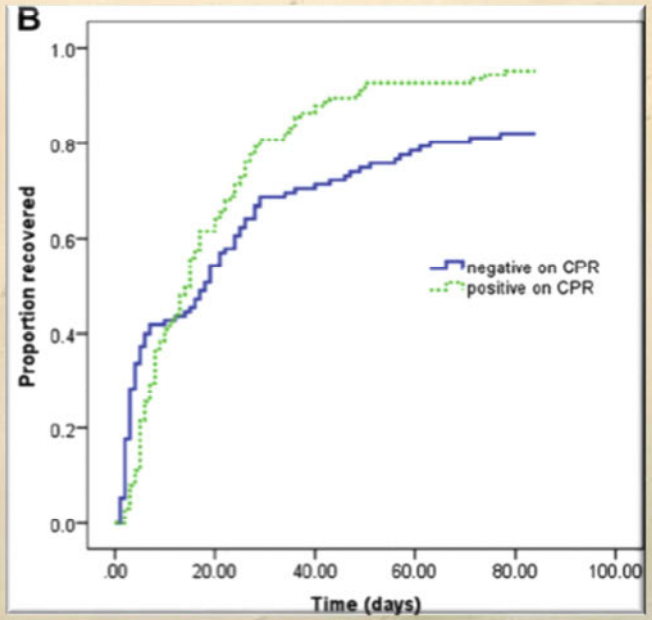
1 - 0.633 = 0.0.367

1 - 0.713 = 0.287

1 - 0.552 = 0.448

Cox Proportional Hazards (Cox Regression)

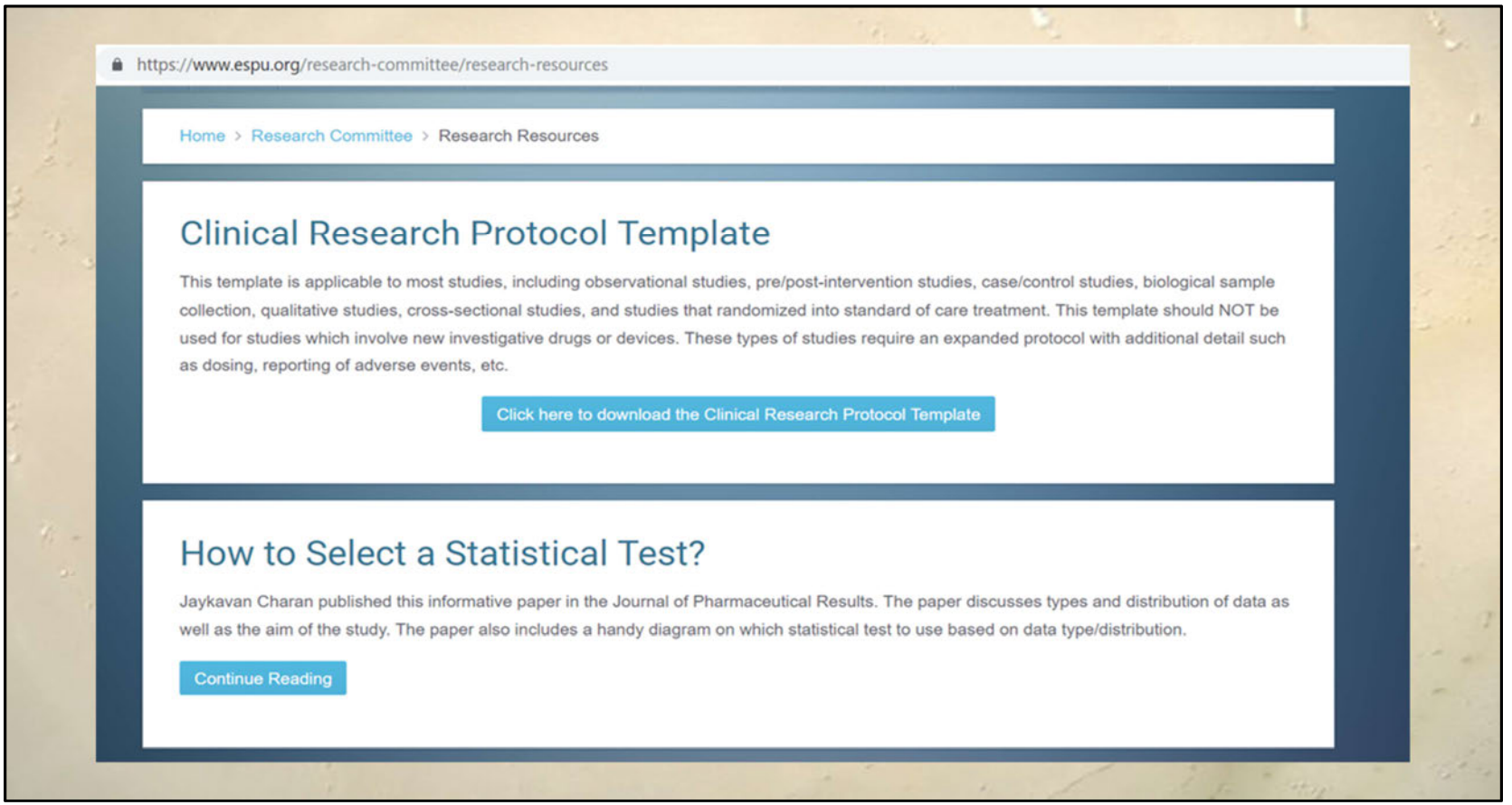
- Estimates hazards ratios (risk)
- Binary dependent variable
- ≥1 independent variables



Cox Regression

Table III. Risk factors for mortality in unadjusted and multivariable model Cox regression analysis.

	Unadjusted			Multivariable Model		
Initial Hospitalization Characteristic	HR	95% CI	p-value	HR	95% CI	p-value
Initial Hospitalization Length of Stay (days)	**1.2**	**1.1 – 1.3**	**<0.0001**	**1.2**	**1.1 – 1.3**	**<0.0001**
Admit Age (days)	1.0	1.0 - 1.0	0.061	1.0	1.0 – 1.0	0.138
<37 week at birth	**5.5**	**2.8 - 11.0**	**<0.0001**	**2.9**	**1.3 – 6.3**	**0.008**
Renal Agenesis	**7.0**	**2.7 – 18.3**	**<0.0001**	2.6	0.9 – 7.4	0.075
Renal Dysplasia	2.1	0.9 – 4.4	0.055	1.0	0.5 – 2.4	0.930
Sepsis	2.4	1.0 – 5.9	0.049	0.8	0.3 – 2.1	0.648
Pulmonary Hypoplasia	**13.2**	**6.7 – 26.3**	**<0.0001**	**7.2**	**3.1 – 16.8**	**<0.0001**
Hospital Volume (reference ≤2 PUV case/yr)						
>2 PUV cases/yr	**0.35**	**0.16 - 0.76**	**0.008**	**0.4**	**0.2 – 0.9**	**0.034**


https://www.espu.org/research-committee/research-resources
Home > Research Committee > Research Resources
Clinical Research Protocol Template
This template is applicable to most studies, including observational studies, pre/post-intervention studies, case/control studies, biological sample collection, qualitative studies, cross-sectional studies, and studies that randomized into standard of care treatment. This template should NOT be used for studies which involve new investigative drugs or devices. These types of studies require an expanded protocol with additional detail such as dosing, reporting of adverse events, etc.
Click here to download the Clinical Research Protocol Template
How to Select a Statistical Test?
Jaykavan Charan published this informative paper in the Journal of Pharmaceutical Results. The paper discusses types and distribution of data as well as the aim of the study. The paper also includes a handy diagram on which statistical test to use based on data type/distribution.
Continue Reading