



Statistics in Healthcare

Unit 7: Overview/Teasers



Overview

- Tests for comparing groups—
unadjusted analyses!

types of outcome data

Outcome Variable	Are the observations independent or correlated? 가 가?		Alternatives (assumptions violated)
	independent	correlated	
Continuous (e.g. pain scale, cognitive function) -	Ttest t ANOVA Linear correlation Linear regression	Paired ttest Repeated-measures ANOVA Mixed models/GEE modeling	Wilcoxon sign-rank test Wilcoxon rank-sum test Kruskal-Wallis test Spearman rank correlation coefficient
Binary or categorical (e.g. fracture yes/no)	Risk difference/Relative risks Chi-square test Logistic regression	McNemar's test Conditional logistic regression GEE modeling	Fisher's exact test McNemar's exact test
Time-to-event (e.g. time to fracture)	Rate ratio Kaplan-Meier statistics Cox regression	Frailty model (beyond the scope of this course)	Time-varying effects (beyond the scope of this course)



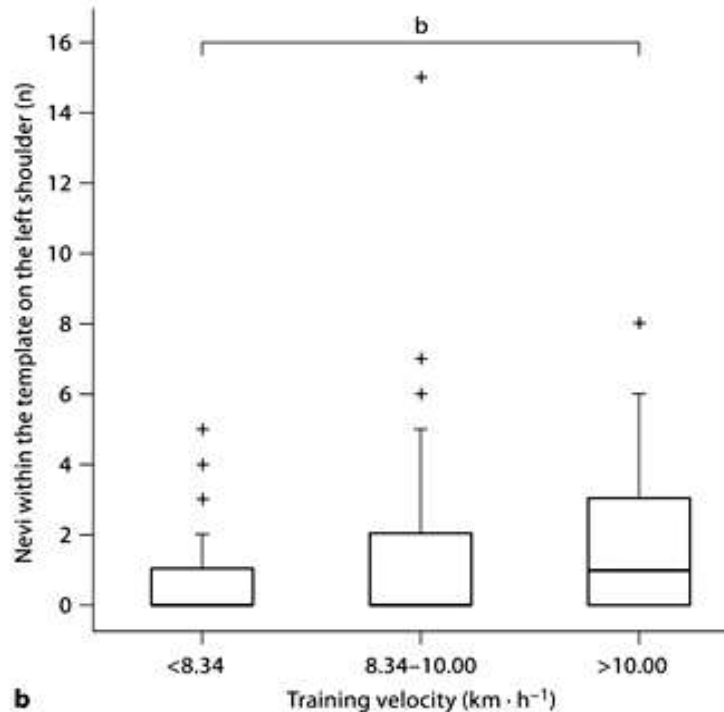
Teaser 1, Unit 7

TABLE 1. Difference between Means of "Before" and "After" Botulinum Toxin A Treatment

	Before BTxnA	After BTxnA
Social skills	5.90	5.84
Academic performance	5.86	5.78
Date success	5.17	5.30
Occupational success	6.08	5.97
Attractiveness	4.94	5.07
Financial success	5.67	5.61
Relationship success	5.68	5.68
Athletic success	5.15	5.38

Reproduced with permission from: DAYAN, S. H., LIEBERMAN, E. D., THAKKAR, N. N., LARIMER, K. A. and ANSTEAD, A. (2008), Botulinum Toxin A Can Positively Impact FirstImpression. Dermatologic Surgery, 34: S40–S47

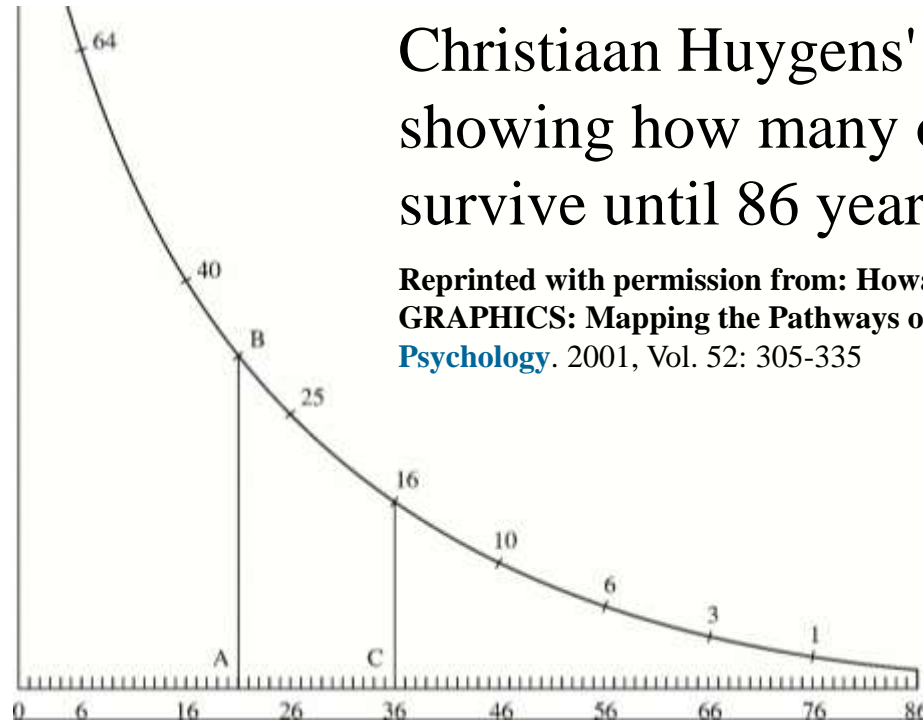
Teaser 2, Unit 7



Study of marathoners and skin lesions (nevi):
Is the number of nevi statistically different in the three training velocity groups?

Reproduced with permission from: Richtig et al. Melanoma Markers in Marathon Runners: Increase with Sun Exposure and Physical Strain. *Dermatology* 2008;217:38-44.

Teaser 3, Unit 7



Christiaan Huygens' 1669 curve
showing how many out of 100 people
survive until 86 years.

Reprinted with permission from: Howard Wainer **STATISTICAL GRAPHICS: Mapping the Pathways of Science.** [Annual Review of Psychology](#). 2001, Vol. 52: 305-335



Statistics in Medicine

Module 1:


Comparing means between 2
groups (or 2 time points)



Assumptions of linear models

Assumptions for linear models (ttest, ANOVA, linear correlation, linear regression):

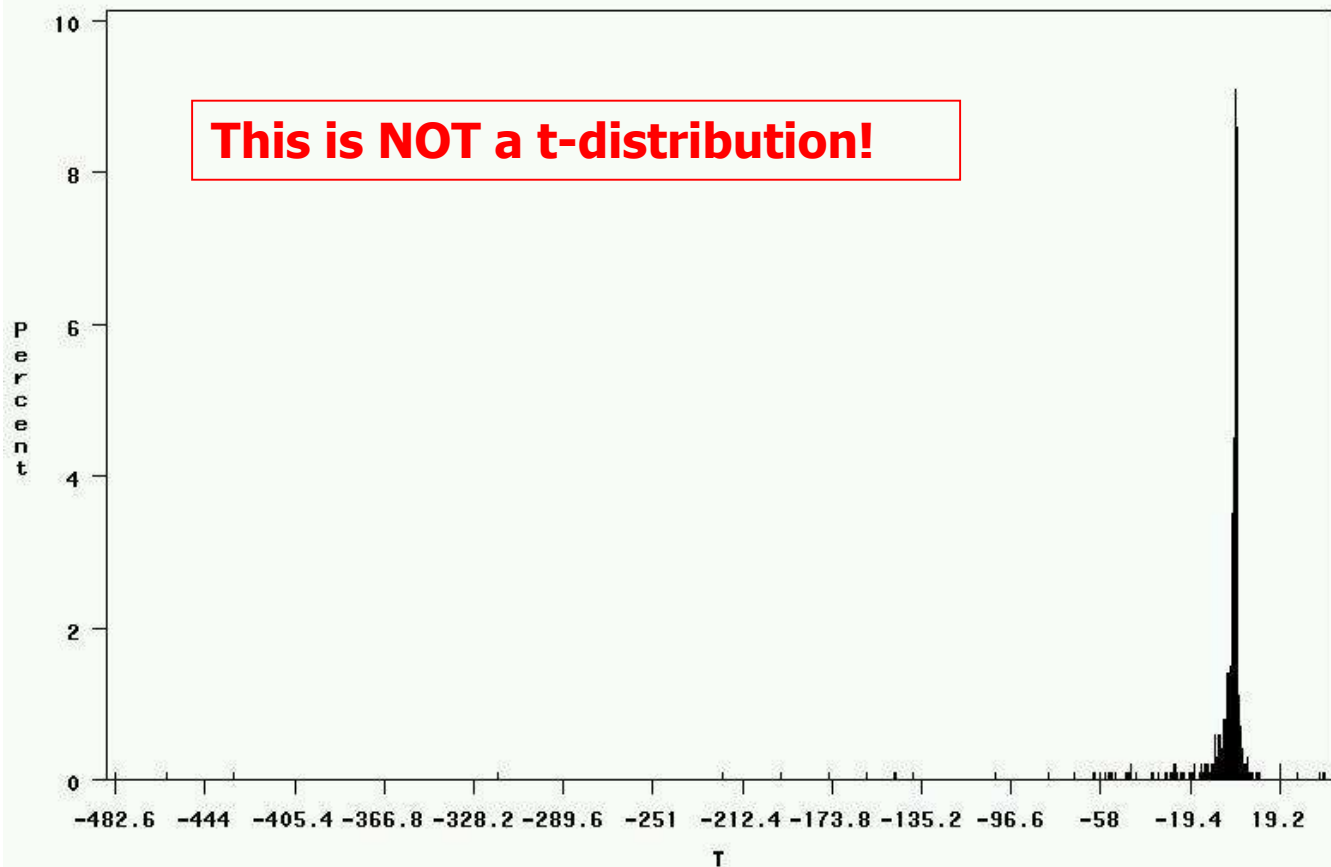
1. Normally distributed outcome variable
 - This assumption is most important for small samples; large samples are quite robust against this assumption because of the central limit theorem (averages are normally distributed even when the underlying trait is not!).
2. Homogeneity of variances
 - Models are robust against this assumption.
 - This assumption is not required for the two-sample ttest if you use the unpooled variance.



Computer simulation: when does the normality assumption matter?

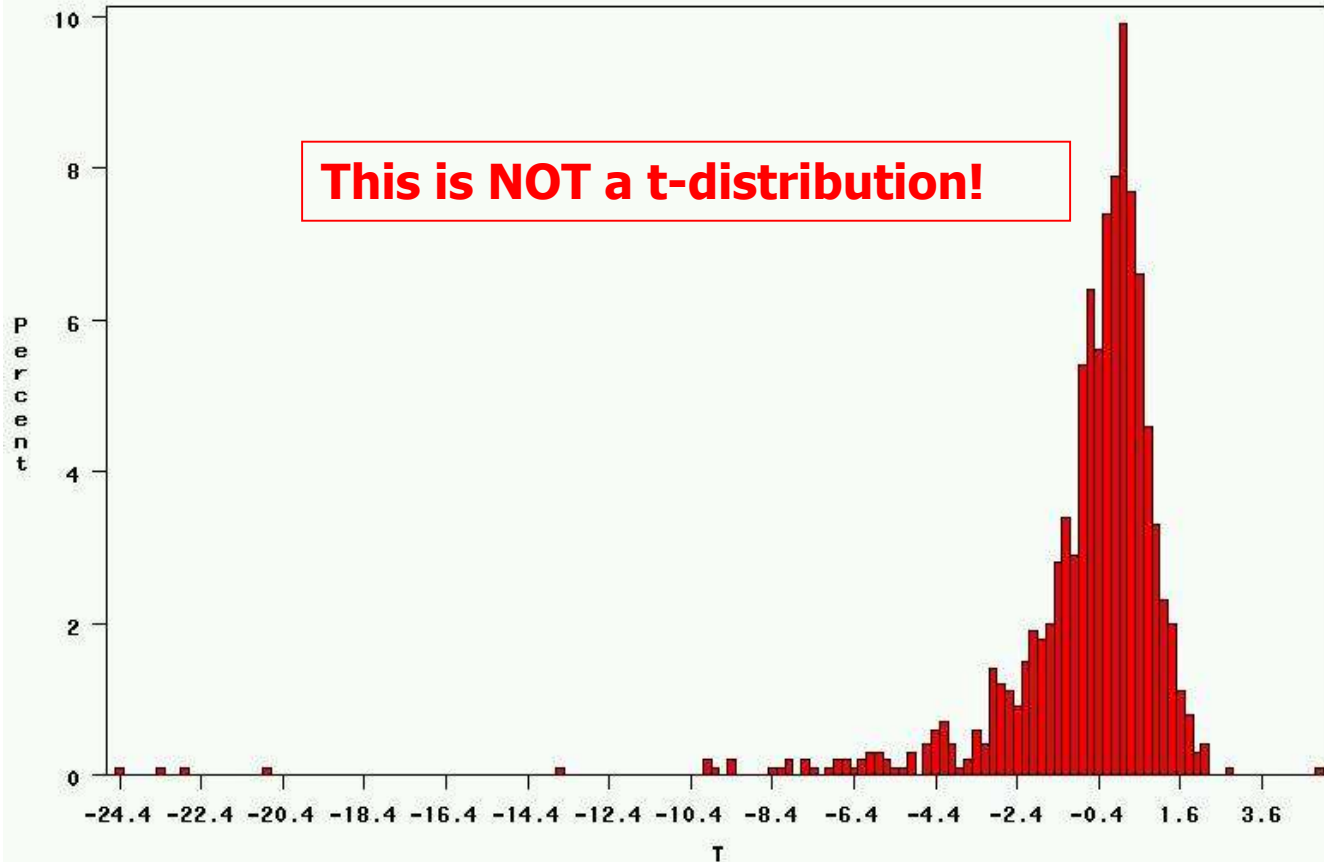
- Computer simulations to observe the distribution of the means when the underlying trait has a highly left-skewed distribution.

$n=2$, underlying distribution is left-skewed (mean=1, SD=1)



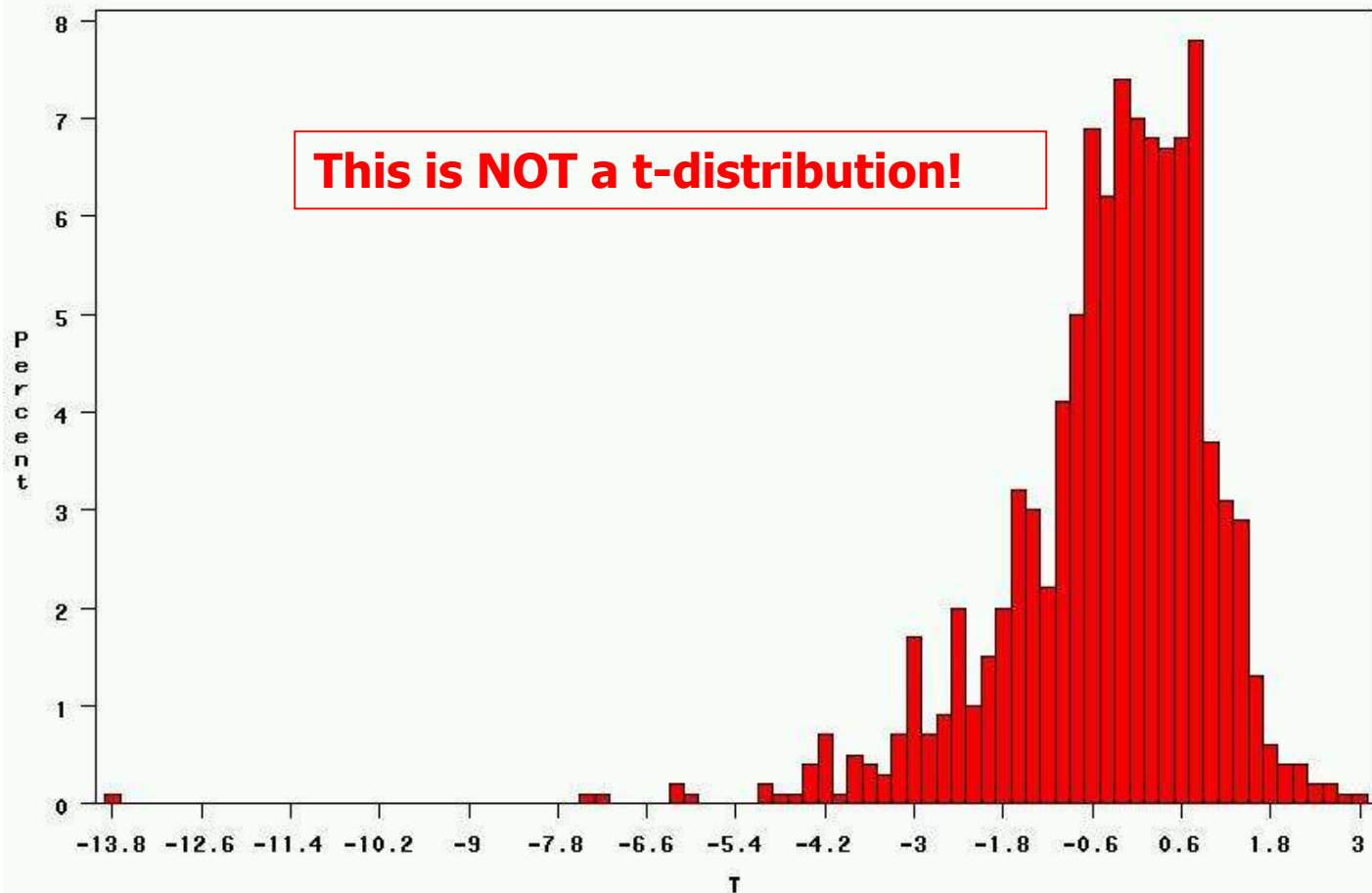
1000 T statistics from averages of 2 from an exponential distribution, unknown sigma

$n=5$



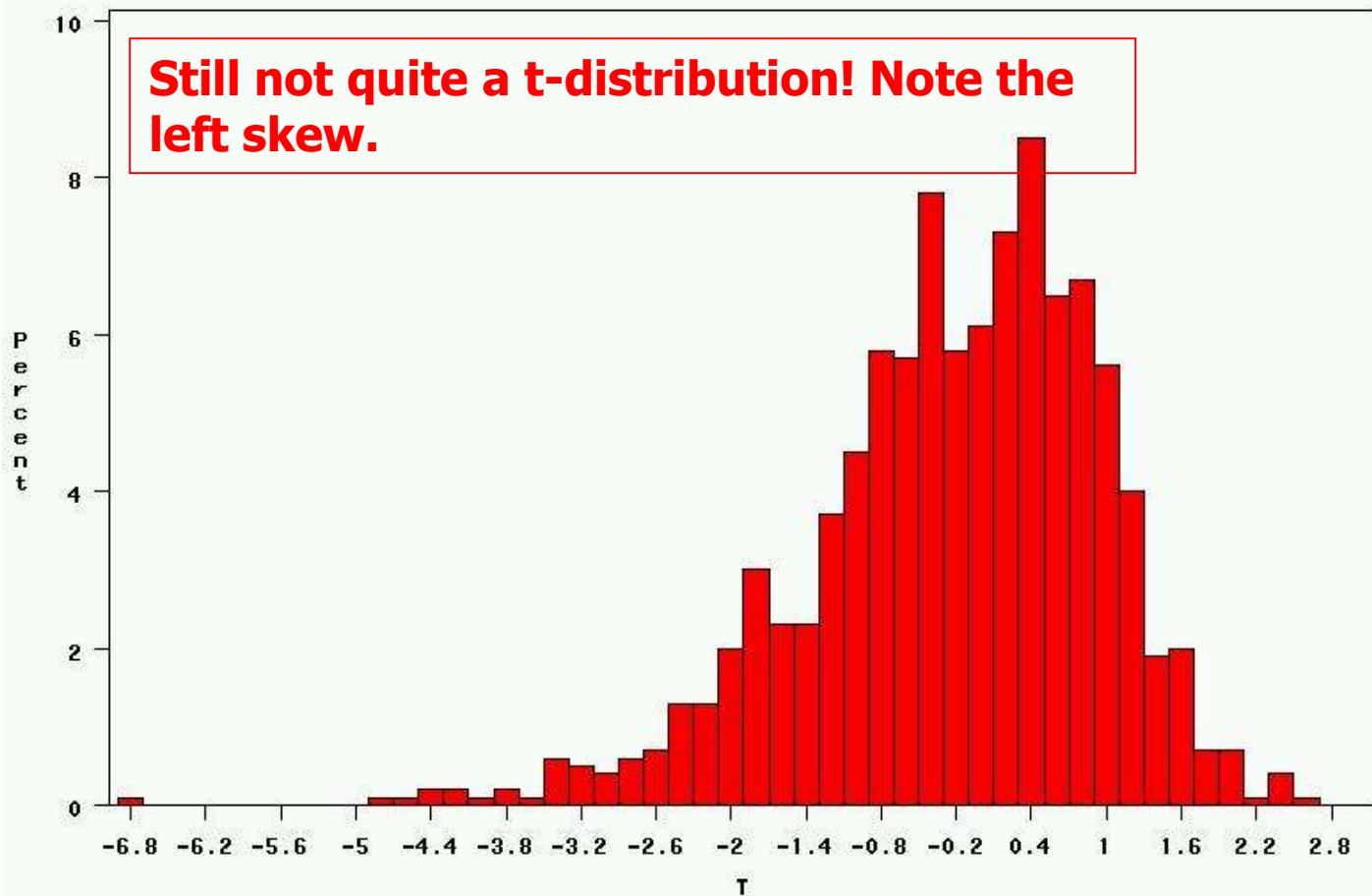
1000 T statistics from averages of 5 from an exponential distribution, unknown sigma

$n=10$



1000 T statistics from averages of 10 from an exponential distribution, unknown sigma

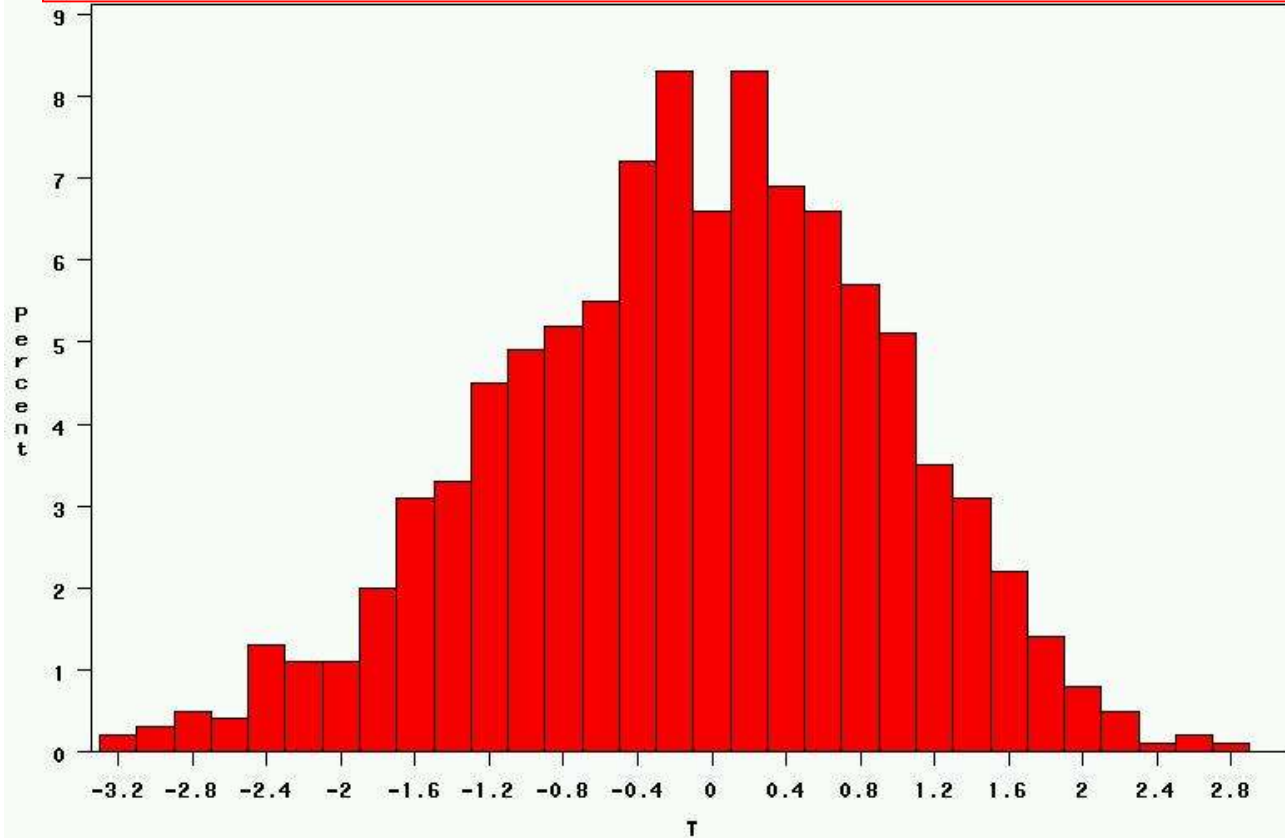
N=30



1000 T statistics from averages of 30 from an exponential distribution, unknown sigma

$N=100$

Now, pretty close to a T-distribution (with 99 degrees of freedom it's also very close to a Z-distribution)!



1000 T statistics from averages of 100 from an exponential distribution, unknown sigma



Conclusions

- If the underlying data are not normally distributed AND n is small**, the means do not follow a t-distribution (so using a ttest will result in erroneous inferences).
- Data transformation or non-parametric tests should be used instead.
- **How small is too small? No hard and fast rule—depends on the true shape of the underlying distribution. Here $N > 30$ (closer to 100) is needed.

Continuous outcome (means)

Outcome Variable	Are the observations independent or correlated?		Alternatives if the normality assumption is violated <u>and</u> small sample size:
	independent	correlated	
Continuous (e.g. pain scale, cognitive function)	Ttest (2 groups) ANOVA (2 or more groups) Pearson's correlation coefficient (1 continuous predictor) Linear regression (multivariate regression technique)	Paired ttest (2 groups or time-points) Repeated-measures ANOVA (2 or more groups or time-points) Mixed models/GEE modeling: (multivariate regression techniques)	<u>Non-parametric statistics</u> Wilcoxon sign-rank test (alternative to the paired ttest) Wilcoxon rank-sum test (alternative to the ttest) Kruskal-Wallis test (alternative to ANOVA) Spearman rank correlation coefficient (alternative to Pearson's correlation coefficient)



Example: two-sample t-test

- In 1980, some researchers reported that “men have more mathematical ability than women” as evidenced by the 1979 SAT’s, where a sample of 30 random male adolescents had a mean score ± 1 standard deviation of 436 ± 77 and 30 random female adolescents scored lower: 416 ± 81 (genders were similar in educational backgrounds, socio-economic status, and age). Is this difference statistically significant?

Continuous outcome (means)

Outcome Variable	Are the observations independent or correlated?		Alternatives if the normality assumption is violated <u>and</u> small sample size:
	independent	correlated	
Continuous (e.g. pain scale, cognitive function)	<p>Ttest (2 groups)</p> <p>ANOVA (2 or more groups)</p> <p>Pearson's correlation coefficient (1 continuous predictor)</p> <p>Linear regression (multivariate regression technique)</p>	<p>Paired ttest (2 groups or time-points)</p> <p>Repeated-measures ANOVA (2 or more groups or time-points)</p> <p>Mixed models/GEE modeling: (multivariate regression techniques)</p>	<p><u>Non-parametric statistics</u></p> <p>Wilcoxon sign-rank test (alternative to the paired ttest)</p> <p>Wilcoxon rank-sum test (alternative to the ttest)</p> <p>Kruskal-Wallis test (alternative to ANOVA)</p> <p>Spearman rank correlation coefficient (alternative to Pearson's correlation coefficient)</p>



Two sample ttest

Statistical question: Is there a difference in SAT math scores between men and women?

- What is the outcome variable? Math SAT scores
 - What type of variable is it? Continuous
 - Is it normally distributed? Yes
 - Are the observations correlated? No
 - Are groups being compared, and if so, how many?
Yes, two
- two-sample ttest



Two-sample ttest mechanics...

- The difference in means follows a T-distribution (Z distribution for larger samples)
 - Assumes that the outcome variable is normally distributed for small n.
- The standard error of the difference in means is (unpooled variance):

$$SE_{(x_1 - x_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$



Two-sample ttest mechanics...

- The standard error of the difference in means is (pooled variance):

$$S_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}$$

where

$$S_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$



Pooled vs. Unpooled variance

Rule of Thumb: Use pooled unless you have a reason not to.

Pooled gives you a more precise estimate of the standard deviation; thus, the T-distribution has more degrees of freedom.

But pooled has an extra assumption: variances are equal between the two groups.

Most statistical programs automatically test this assumption for you ("Equality of Variances" test). If $p < .05$, this suggests unequal variances, and better to use unpooled variance ttest.



Data Summary

	n	Sample Mean	Sample Standard Deviation
Group 1: women	30	416	81
Group 2: men	30	436	77



Two-sample t-test

1. Define your hypotheses (null, alternative)

$$H_0: \text{♂-♀ math SAT} = 0$$

$$H_1: \text{♂-♀ math SAT} \neq 0 \text{ [two-sided]}$$



Two-sample t-test

2. Specify your null distribution:

F and M have approximately equal standard deviations/variances, so make a “pooled” estimate of standard deviation/variance:

$$s_p = 79.02 \quad (\text{If } S_1 \sim S_2, s_p = \frac{81 + 77}{2} = 79)$$

The standard error of a difference of two means is:

$$\sqrt{\frac{s_p^2}{n} + \frac{s_p^2}{m}} = \sqrt{\frac{79^2}{30} + \frac{79^2}{30}} = 20.4$$

Differences in means follow a T-distribution for small samples; Z-distribution for large samples...



Two-sample t-test

3. Observed difference in our experiment = 20 points



Two-sample t-test

4. Calculate the p-value of what you observed

$$T_{58} = \frac{20 - 0}{20.4} = .98$$

$$p = .33$$

5. Do not reject null! No evidence that men are better in math ;)



T-value to p-value calculator...

Describe the random variable	<input type="text" value="t score"/>
Degrees of freedom	<input type="text" value="58"/>
t score	<input type="text" value=".98"/>
Cumulative probability: $P(T \leq .98)$	<input type="text" value="0.8344"/>

<http://stattrek.com/online-calculator/t-distribution.aspx>



T-value for 95% CI...

Describe the random variable	<input type="text" value="t score"/>
Degrees of freedom	<input type="text" value="58"/>
t score	<input type="text" value="-2.002"/>
Cumulative probability: $P(T \leq t)$	<input type="text" value=".025"/>
<input type="button" value="Calculate"/>	

<http://stattrek.com/online-calculator/t-distribution.aspx>

Corresponding confidence interval...



$$20 \pm 2.00 * 20.4 = (-20.8) \text{ to } (60.8)$$

Note that the 95% confidence interval crosses 0 (the null value).

Continuous outcome (means)

Outcome Variable	Are the observations independent or correlated?		Alternatives if the normality assumption is violated <u>and</u> small sample size:
	independent	correlated	
Continuous (e.g. pain scale, cognitive function)	<p>Ttest (2 groups)</p> <p>ANOVA (2 or more groups)</p> <p>Pearson's correlation coefficient (1 continuous predictor)</p> <p>Linear regression (multivariate regression technique)</p>	<p>Paired ttest (2 groups or time-points)</p> <p>Repeated-measures ANOVA (2 or more groups or time-points)</p> <p>Mixed models/GEE modeling: (multivariate regression techniques)</p>	<p><u>Non-parametric statistics</u></p> <p>Wilcoxon sign-rank test (alternative to the paired ttest)</p> <p>Wilcoxon rank-sum test (alternative to the ttest)</p> <p>Kruskal-Wallis test (alternative to ANOVA)</p> <p>Spearman rank correlation coefficient (alternative to Pearson's correlation coefficient)</p>



Example: paired ttest

TABLE 1. Difference between Means of "Before" and "After" Botulinum Toxin A Treatment

	Before BTxnA	After BTxnA	Difference	Significance
Social skills	5.90	5.84	NS	.293
Academic performance	5.86	5.78	.08	.068
Date success	5.17	5.30	.13	.014*
Occupational success	6.08	5.97	.11	.013*
Attractiveness	4.94	5.07	.13	.030*
Financial success	5.67	5.61	NS	.230
Relationship success	5.68	5.68	NS	.967
Athletic success	5.15	5.38	.23	.000**

* Significant at 5% level.

** Significant at 1% level.

Reproduced with permission from: DAYAN, S. H., LIEBERMAN, E. D., THAKKAR, N. N., LARIMER, K. A. and ANSTEAD, A. (2008), Botulinum Toxin A Can Positively Impact FirstImpression. Dermatologic Surgery, 34: S40–S47



Paired ttest

Statistical question: Is there a difference in date success after BoTox?

- What is the outcome variable? Date success
- What type of variable is it? Continuous
- Is it normally distributed? Yes
- Are the observations correlated? Yes, it's the same patients before and after
- How many time points are being compared? Two
→ paired ttest



Paired ttest mechanics

1. Calculate the change in date success score for each person.
2. Calculate the average change in date success for the sample.
(=.13)
3. Calculate the standard error of the change in date success.
(=.05)
4. Calculate a T-statistic by dividing the mean change by the standard error ($T = .13 / .05 = 2.6$).
5. Look up the corresponding p-values. ($T = 2.6$ corresponds to $p = .014$).
6. Significant p-values indicate that the average change is significantly different than 0.



Paired ttest example 2...

Patient	BP Before (diastolic)	BP After
1	100	92
2	89	84
3	83	80
4	98	93
5	108	98
6	95	90



Example problem: paired ttest

Patient	Diastolic BP Before	D. BP After	Change
1	100	92	-8
2	89	84	-5
3	83	80	-3
4	98	93	-5
5	108	98	-10
6	95	90	-5

Null Hypothesis: Average Change = 0



Example problem: paired ttest

$$\bar{X} = \frac{-8 - 5 - 3 - 5 - 10 - 5}{6} = \frac{-36}{6} = -6$$

$$s_x = \sqrt{\frac{(-8 - -6)^2 + (-5 - -6)^2 + (-3 - -6)^2 + \dots}{5}} =$$
$$\sqrt{\frac{4 + 1 + 9 + 1 + 16 + 1}{5}} = \sqrt{\frac{32}{5}} = 2.5$$

$$s_{\bar{x}} = \frac{2.5}{\sqrt{6}} = 1.0$$

$$T_5 = \frac{-6 - 0}{1.0} = -6$$

Null Hypothesis: Average Change = 0

**With 5 df, T=-6
corresponds to p=.0018**

Change

-8

-5


-3

-5

-10

-5


Online tools for finding T-distribution probabilities...



Describe the random variable	t score	
Degrees of freedom	5	
t score	-6	
Cumulative probability: $P(T \leq -6)$	0.0009	
		Calculate

<http://stattrek.com/online-calculator/t-distribution.aspx>

Find the T-value for 95% confidence...



Describe the random variable

Degrees of freedom

t score

Cumulative probability: $P(T \leq t)$

<http://stattrek.com/online-calculator/t-distribution.aspx>



Example problem: paired ttest

$$\begin{aligned} & \mathbf{95\% \text{ CI} : - 6 \pm 2.571 * (1.0)} \\ & \mathbf{= (-3.43 , - 8.571)} \end{aligned}$$

Note: does not include 0.

Change
-8
-5
-3
-5
-10
-5

Use the paired ttest to compare correlated samples!

Twin pair	Diastolic blood pressure in the less active twin (mmHg)	Diastolic blood pressure in the more active twin (mmHg)	Difference (more active – less active) (mmHg)
1	87	82	-5
2	88	83	-5
3	80	78	-2
4	79	80	+1
5	77	71	-6
6	69	65	-4
Mean (SD)	80.0 (7.0)	76.5 (7.1)	-3.5 (2.6)
Test statistic	<p><u>Two-sample ttest (incorrect analysis):</u></p> $T_{10} = \frac{-3.5}{\sqrt{\frac{7.0^2}{6} + \frac{7.0^2}{6}}} = -0.86$ <p>$p = .41$</p>		<p><u>Paired ttest (correct analysis):</u></p> $T_5 = \frac{-3.5}{\sqrt{\frac{2.6^2}{6}}} = -3.31$ <p>$p = .02$</p>



Statistics in Medicine

Module 2:

Comparing means between more
than 2 groups (or time points)

Continuous outcome (means)

Outcome Variable	Are the observations independent or correlated?		Alternatives if the normality assumption is violated <u>and</u> small sample size:
	independent	correlated	
Continuous (e.g. pain scale, cognitive function)	Ttest (2 groups) ANOVA (2 or more groups) Pearson's correlation coefficient (1 continuous predictor) Linear regression (multivariate regression technique)	Paired ttest (2 groups or time-points) Repeated-measures ANOVA (2 or more groups or time-points) Mixed models/GEE modeling: (multivariate regression techniques)	<u>Non-parametric statistics</u> Wilcoxon sign-rank test (alternative to the paired ttest) Wilcoxon rank-sum test (alternative to the ttest) Kruskal-Wallis test (alternative to ANOVA) Spearman rank correlation coefficient (alternative to Pearson's correlation coefficient)

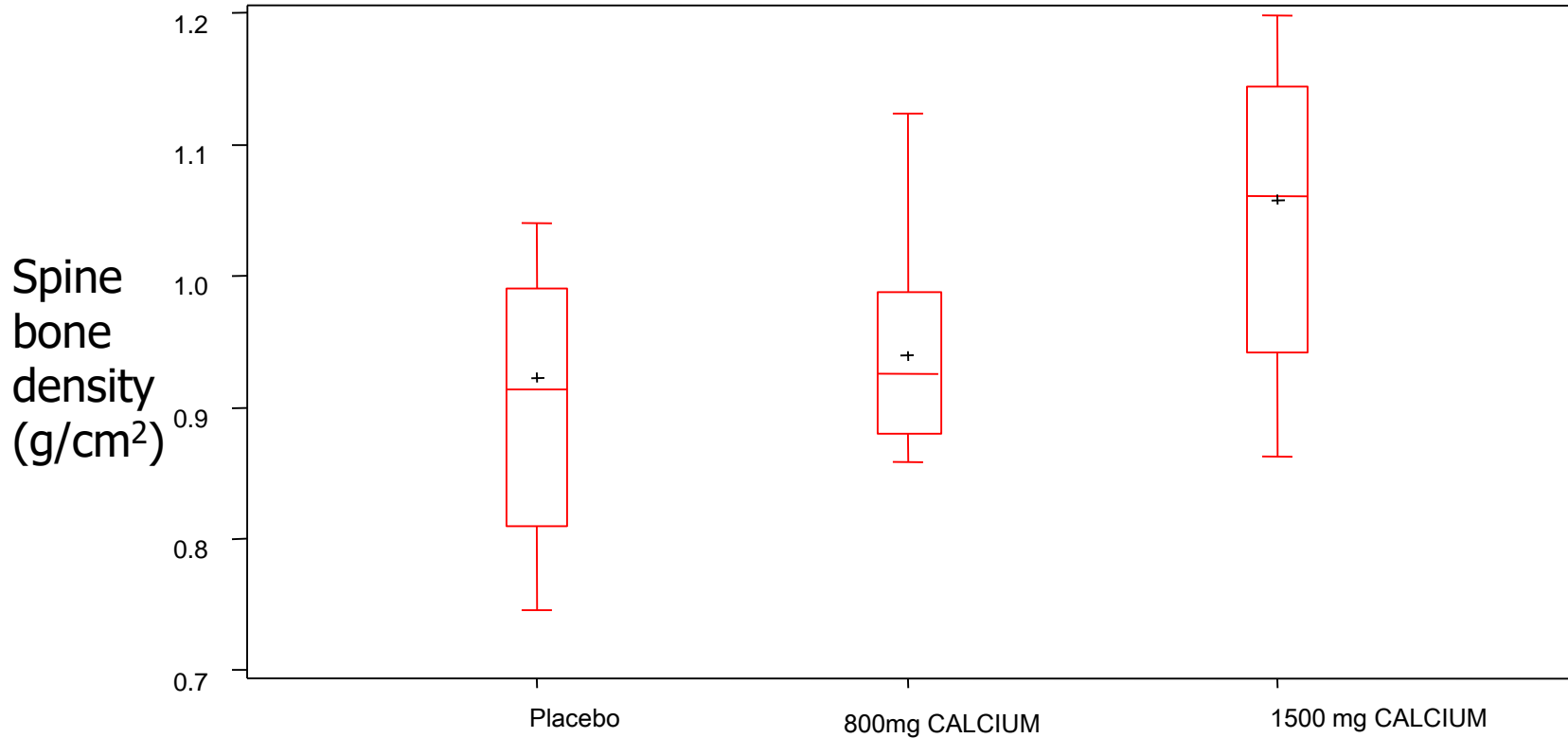


Example

Hypothetical trial: Randomize 33 subjects to three groups: 800 mg calcium supplement vs. 1500 mg calcium supplement vs. placebo.

Compare the spine bone density of all 3 groups after 1 year.

Results: spine bone density at year 1





ANOVA

Statistical question: Is there a difference in final spine bone density in the three treatment groups?

- What is the outcome variable? Spine bone density
- What type of variable is it? Continuous
- Is it normally distributed? Yes (need to test from data)
- Are the observations correlated? No
- Are groups being compared and, if so, how many? Yes, three

→ ANOVA



Analysis of Variance

- Assumptions, same as ttest
 - Normally distributed outcome
 - Equal variances between the groups
 - Groups are independent



Question: *Why not just do 3 pairwise ttests?*

Answer: because, at a type I error rate of 5% each test, this means you have an overall chance of up to $1 - (.95)^3 = 14\%$ of making a type-I error

(this is the type I error if the comparisons are independent, which they are not)



Hypotheses of One-Way ANOVA (Global test!)

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \dots$$

H_a : Not all of the population means are the same

[Alt. hypothesis: at least one of the mean is different (group unknown)]



ANOVA

- It's like this: If I have three groups to compare:
 - I could do three pair-wise ttests, but this would increase my type I error
 - So, instead I want to look at the pairwise differences "all at once."
 - To do this, I can recognize that variance is a statistic that let's me look at more than one difference at a time...



The “F-test”

Is the difference in the means of the groups more than background noise (=variability within groups)?

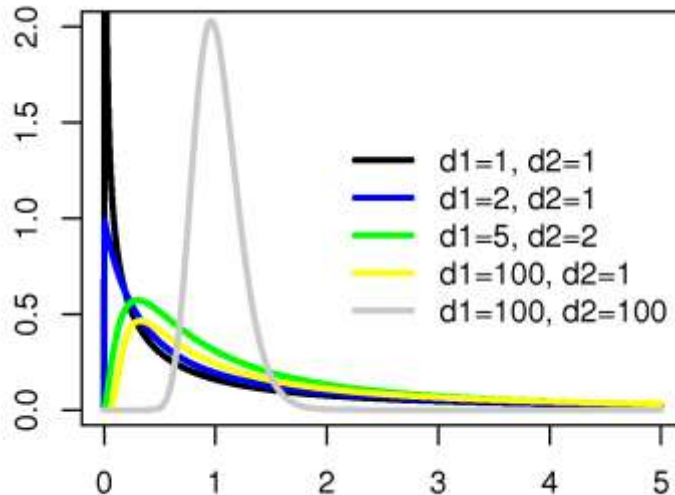
Summarizes the mean differences between all groups at once.

$$F = \frac{\text{Variability between groups}}{\text{Variability within groups}}$$

Analogous to pooled variance from a ttest.

The F-distribution

- The F-distribution is a continuous probability distribution that depends on two parameters n and m (numerator and denominator degrees of freedom, respectively):





The F-distribution

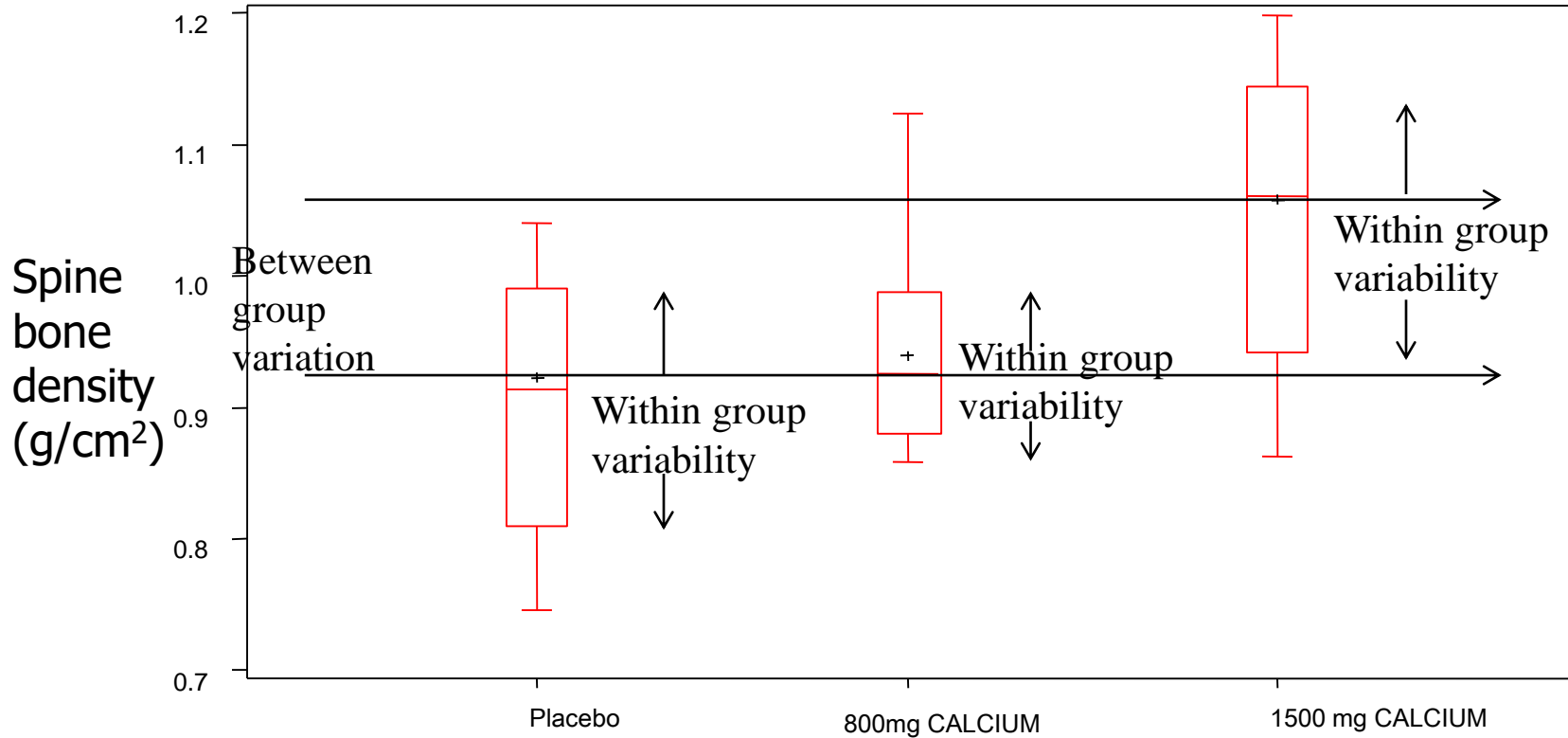
- A ratio of variances follows an F-distribution:

$$\frac{\sigma_{between}^2}{\sigma_{within}^2} \sim F_{n, d}$$

- The F-test tests the hypothesis that two variances are equal.
- F will be close to 1 if sample variances are equal.

$$\begin{aligned} H_0 : \sigma_{between}^2 &= \sigma_{within}^2 \\ H_a : \sigma_{between}^2 &\neq \sigma_{within}^2 \end{aligned}$$

Results: spine bone density at year 1



Group means and standard deviations



- Placebo group (n=11):
 - Mean spine BMD = .92 g/cm²
 - standard deviation = .10 g/cm²
- 800 mg calcium supplement group (n=11)
 - Mean spine BMD = .94 g/cm²
 - standard deviation = .08 g/cm²
- 1500 mg calcium supplement group (n=11)
 - Mean spine BMD = 1.06 g/cm²
 - standard deviation = .11 g/cm²

Between-group
variation.

The size of the
groups.

The difference of
each group's
mean from the
overall mean.

The F-Test

$$s_{between}^2 = ns_{\bar{x}}^2 = 11 * \left(\frac{(.92 - .97)^2 + (.94 - .97)^2 + (1.06 - .97)^2}{3 - 1} \right) = .063$$

$$s_{within}^2 = avg s^2 = \frac{1}{3} (.10^2 + .08^2 + .11^2) = .0095$$

$$F_{2,30} = \frac{s_{between}^2}{s_{within}^2} = \frac{.063}{.0095} = 6.6$$

The average
amount of
variation within
groups.

Each group's variance.

Large F value indicates
that the between group
variation exceeds the
within group variation
(=the background noise).



ANOVA summary

- A statistically significant ANOVA (F-test) only tells you that *at least* two of the groups differ, but not which ones differ.
- Determining *which* groups differ (when it's unclear) requires more sophisticated analyses to correct for the problem of multiple comparisons...



Correction for multiple comparisons

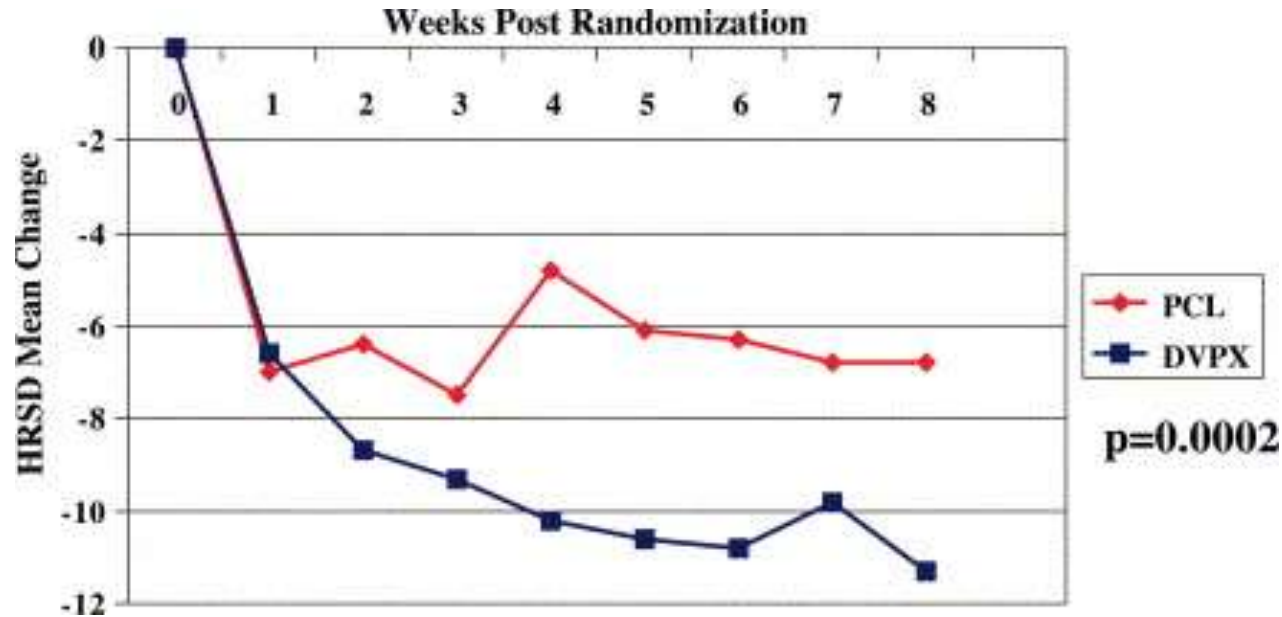
How to correct for multiple comparisons *post-hoc*...

- Bonferroni correction
- Holm/Hochberg
- Tukey (adjusts p)
- Scheffe (adjusts p)

Continuous outcome (means)

Outcome Variable	Are the observations independent or correlated?		Alternatives if the normality assumption is violated <u>and</u> small sample size:
	independent	correlated	
Continuous (e.g. pain scale, cognitive function)	<p>Ttest (2 groups)</p> <p>ANOVA (2 or more groups)</p> <p>Pearson's correlation coefficient (1 continuous predictor)</p> <p>Linear regression (multivariate regression technique)</p>	<p>Paired ttest (2 groups or time-points)</p> <p>Repeated-measures ANOVA (2 or more groups or time-points)</p> <p>Mixed models/GEE modeling: (multivariate regression techniques)</p>	<p><u>Non-parametric statistics</u></p> <p>Wilcoxon sign-rank test (alternative to the paired ttest)</p> <p>Wilcoxon rank-sum test (alternative to the ttest)</p> <p>Kruskal-Wallis test (alternative to ANOVA)</p> <p>Spearman rank correlation coefficient (alternative to Pearson's correlation coefficient)</p>

Divalproex vs. placebo for treating bipolar depression



Reproduced with permission from: Davis et al. "Divalproex in the treatment of bipolar depression: A placebo controlled study." *J Affective Disorders* 85 (2005) 259-266.



Repeated-measures ANOVA

Statistical question: Do subjects in the treatment group have greater reductions in depression scores over time than those in the control group?

- What is the outcome variable? Depression score
 - What type of variable is it? Continuous
 - Is it normally distributed? Yes
 - Are the observations correlated? Yes, there are multiple measurements on each person
 - How many time points are being compared? >2
- repeated-measures ANOVA



Repeated-measures ANOVA

- For before and after studies, a paired ttest will suffice.
- For more than two time periods, you need repeated-measures ANOVA.
- Serial paired ttests is incorrect, because this strategy will increase your type I error.



Repeated-measures ANOVA

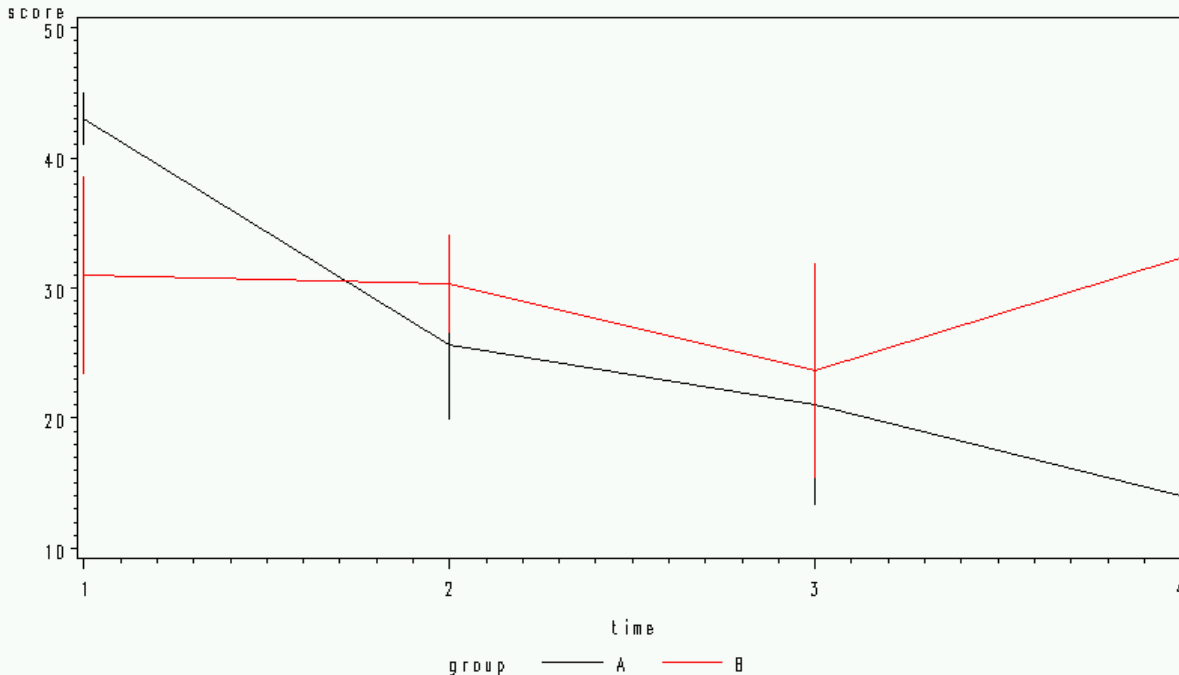
- Answers the following questions, taking into account the fact the correlation within subjects:
 - Are there significant differences across time periods?
 - Are there significant differences between groups (=your categorical predictor)?
 - Are there significant differences between groups in their *changes over time*?



Repeated-measures ANOVA...

- Overall, are there significant differences between time points?
 - Time factor
- Do the two groups differ at any time points?
 - Group factor
- Do the two groups differ in their responses over time?**
 - Group x time factor

Repeated-measures ANOVA

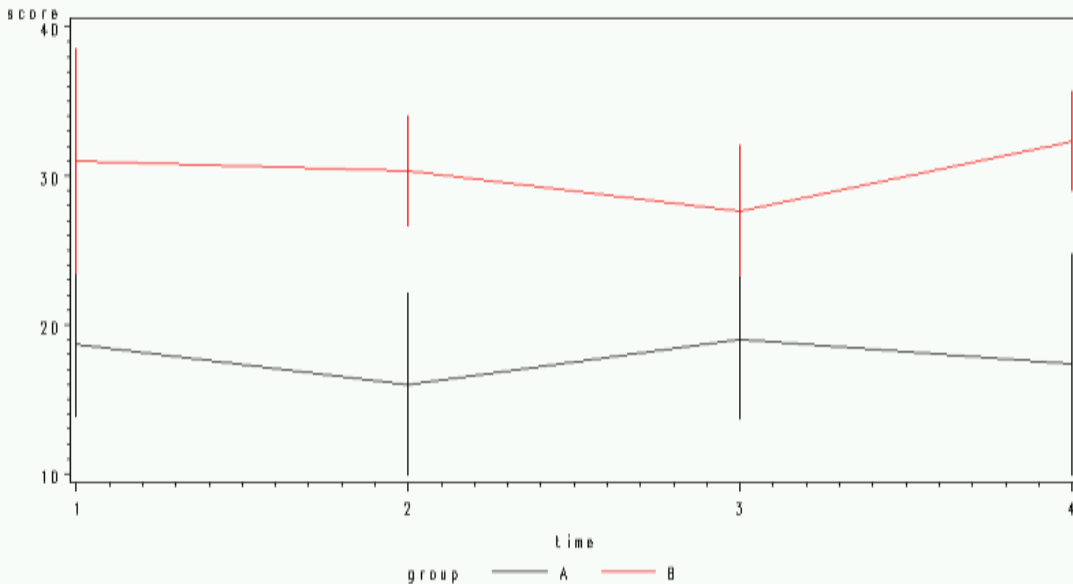


Time is significant.

Group*time is significant.

Group is not significant.

Repeated-measures ANOVA

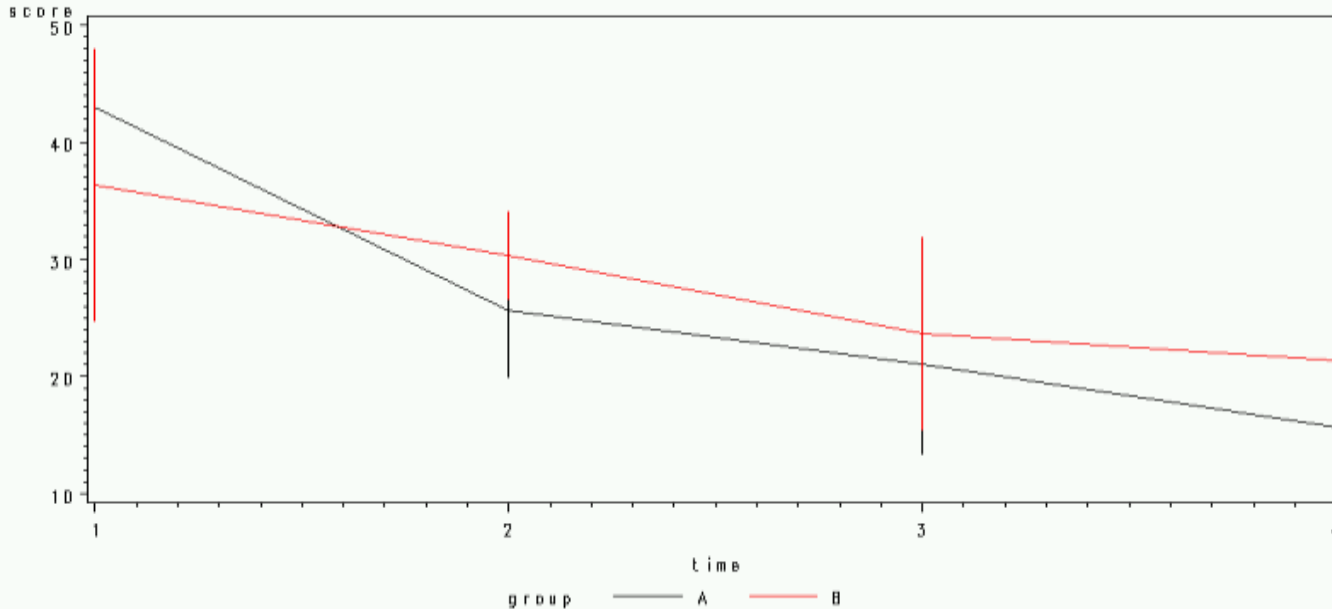


Time is not significant.

Group*time is not significant.

Group IS significant.

Repeated-measures ANOVA



Time is significant.

Group is not significant.

Time*group is not significant.



Statistics in Medicine

Module 3:

Alternative tests to the ttest and
ANOVA (non-parametric tests)

Continuous outcome (means)

Outcome Variable	Are the observations independent or correlated?		Alternatives if the normality assumption is violated <u>and</u> small sample size:
	independent	correlated	
Continuous (e.g. pain scale, cognitive function)	<p>Ttest (2 groups)</p> <p>ANOVA (2 or more groups)</p> <p>Pearson's correlation coefficient (1 continuous predictor)</p> <p>Linear regression (multivariate regression technique)</p>	<p>Paired ttest (2 groups or time-points)</p> <p>Repeated-measures ANOVA (2 or more groups or time-points)</p> <p>Mixed models/GEE modeling: (multivariate regression techniques)</p>	<p>Non-parametric statistics</p> <p>Wilcoxon sign-rank test (alternative to the paired ttest)</p> <p>Wilcoxon rank-sum test (alternative to the ttest)</p> <p>Kruskal-Wallis test (alternative to ANOVA)</p> <p>Spearman rank correlation coefficient (alternative to Pearson's correlation coefficient)</p>

Recall: hypothetical weight loss trial...



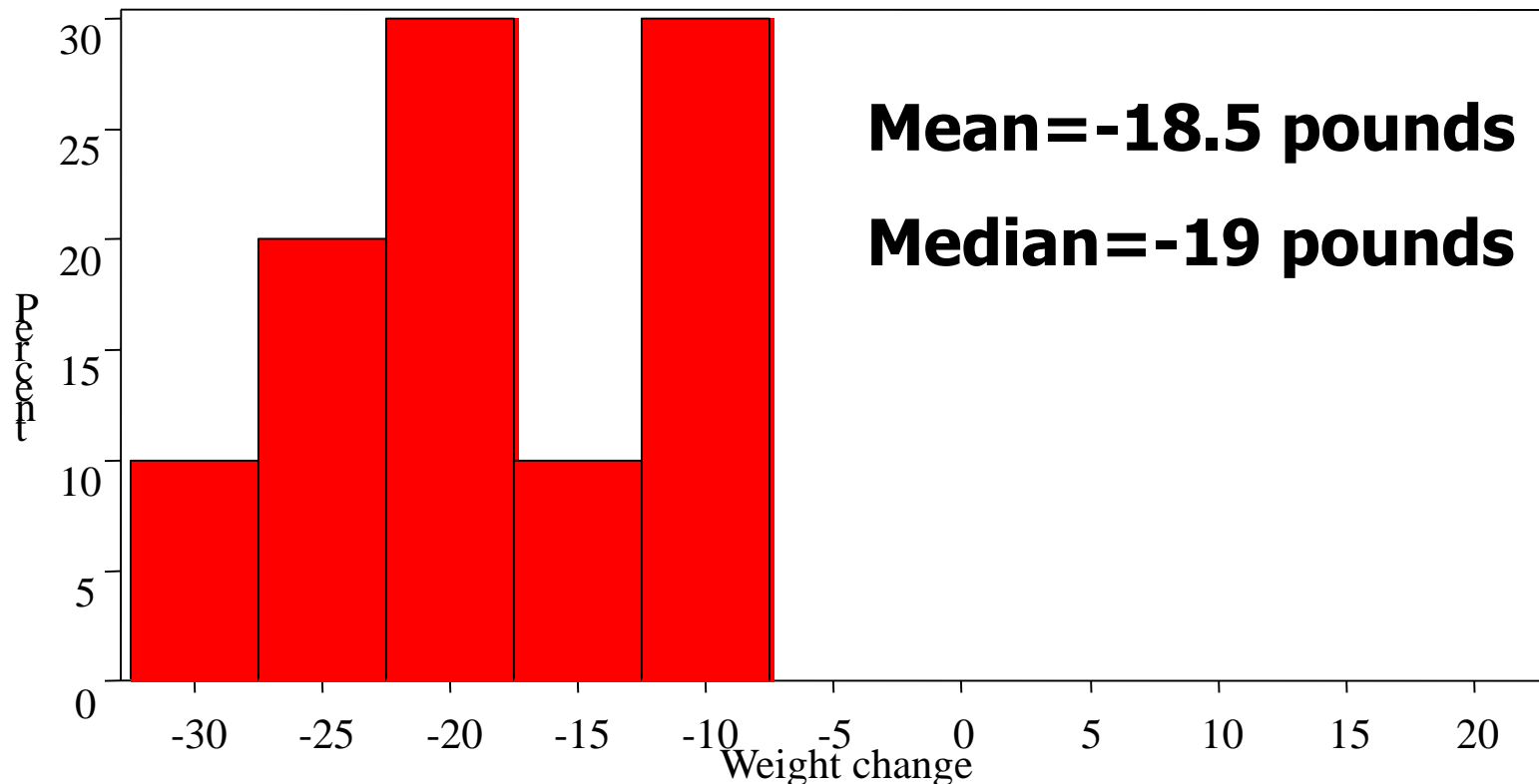
10 dieters following diet 1 vs. 10 dieters following diet 2

Group 1 ($n=10$) loses an average of 34.5 lbs.

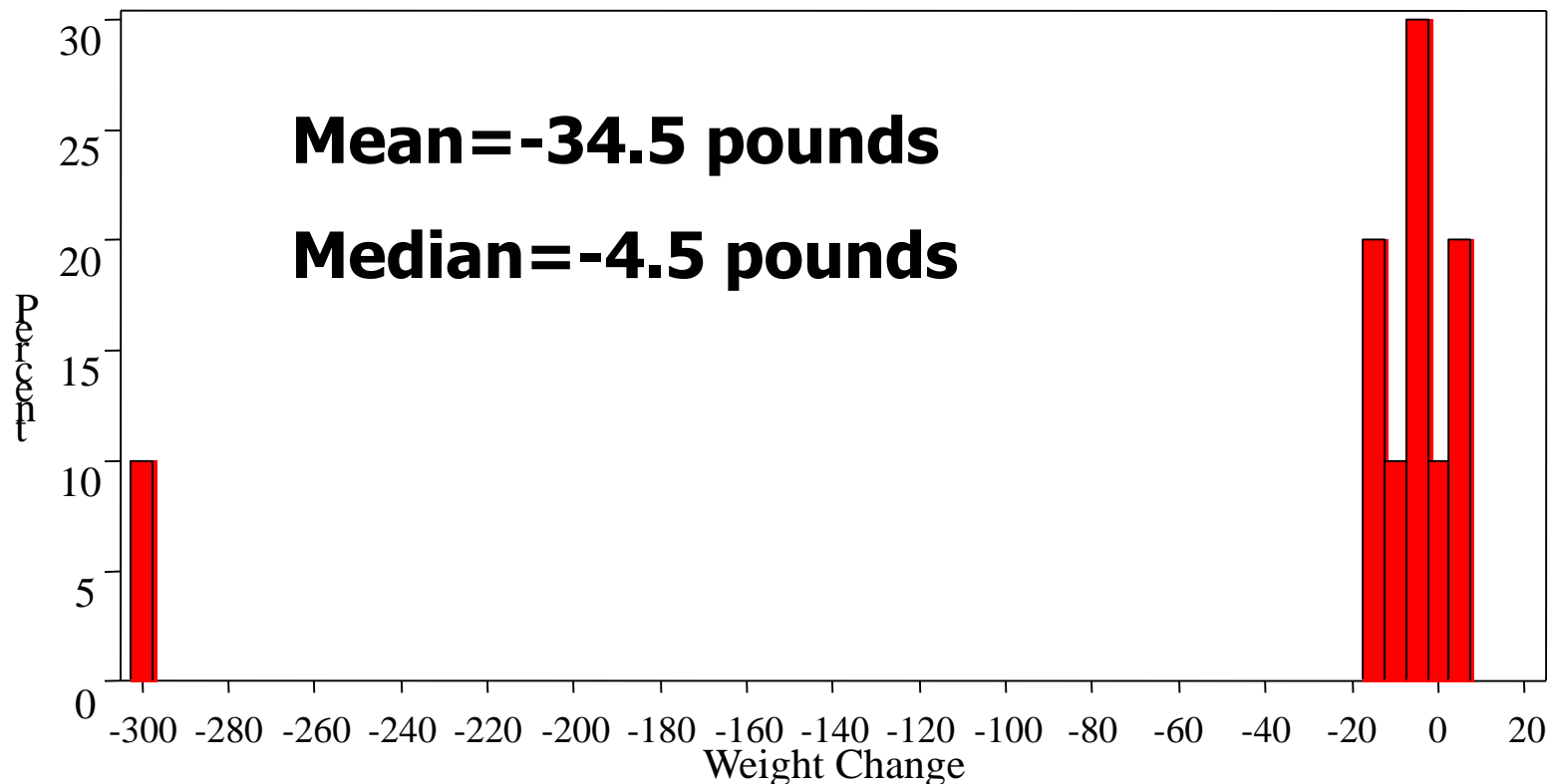
Group 2 ($n=10$) loses an average of 18.5 lbs.

Conclusion: diet 1 is better?

Histogram, diet 2...



Histogram, diet 1...





The data...

Diet 1, change in weight (lbs):

+4, +3, 0, -3, -4, -5, -11, -14, -15, -300

Diet 2, change in weight (lbs)

-8, -10, -12, -16, -18, -20, -21, -24, -26, -30



Wilcoxon rank-sum (Mann-Whitney U) test

Statistical question: Is there a difference in weight loss between the two diets? (Diet2 lose more weight?)

- What is the outcome variable? Weight change
 - What type of variable is it? Continuous
 - Is it normally distributed? **No** (and small n)
 - Are the observations correlated? No
 - Are groups being compared, and if so, how many? two
- Wilcoxon rank-sum test (equivalent to the Mann-Whitney U test!)



Rank the data (ignoring groups)...

Diet 1, change in weight (lbs):

+4, +3, 0, -3, -4, -5, -11, -14, -15, -300

Ranks: 1 2 3 4 5 6 9 11 12 20

Diet 2, change in weight (lbs)

-8, -10, -12, -16, -18, -20, -21, -24, -26, -30

Ranks: 7 8 10 13 14 15 16 17 18 19



Sum the ranks...

**Wilcoxon rank-sum test
compares these numbers
accounting for any
differences in the sample
sizes of the two groups.**

Diet 1, change in weight (lbs):

+4, +3, 0, -3, -4, -5, -11, -14, -15, -300

Ranks: 1 2 3 4 5 6 9 11 12 20

Sum of the ranks: $1+2+3+4+5+6+9+11+12+20 = 73$

**Diet 2 is
superior to
Diet 1, $p=.009$
(single sided).**

Diet 2, change in weight (lbs)

-8, -10, -12, -16, -18, -20, -21, -24, -26, -30

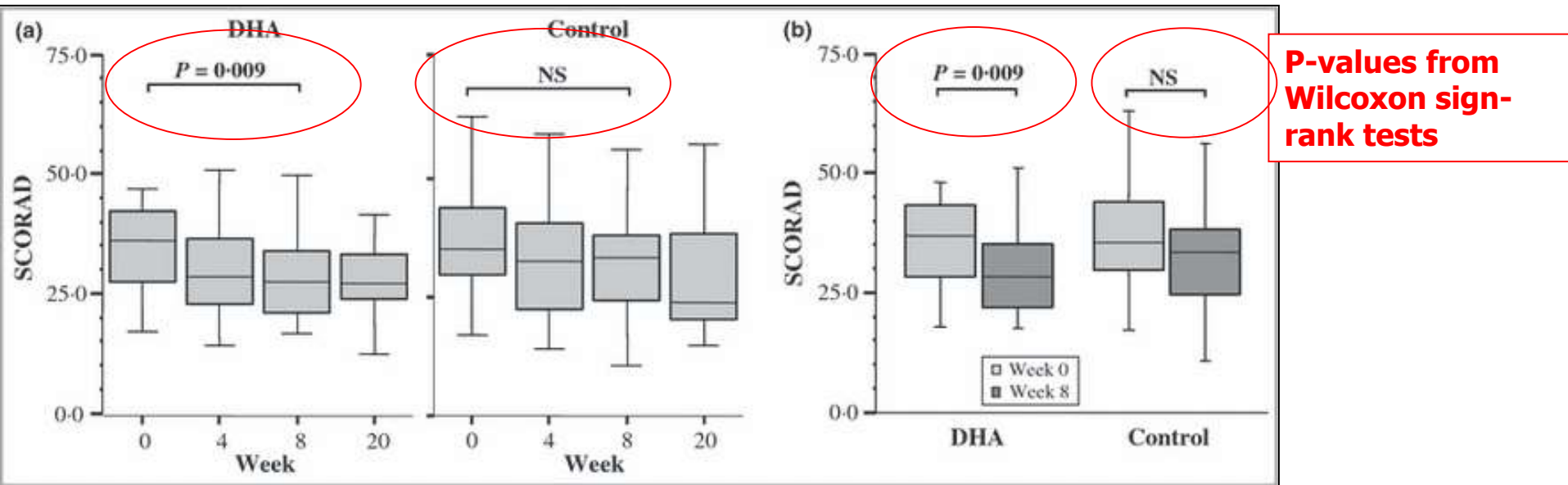
Ranks: 7 8 10 13 14 15 16 17 18 19

Sum of the ranks: $7+8+10+13+14+15+16+17+18+19 = 137$

Continuous outcome (means)

Outcome Variable	Are the observations independent or correlated?		Alternatives if the normality assumption is violated <u>and</u> small sample size:
	independent	correlated	
Continuous (e.g. pain scale, cognitive function)	<p>Ttest (2 groups)</p> <p>ANOVA (2 or more groups)</p> <p>Pearson's correlation coefficient (1 continuous predictor)</p> <p>Linear regression (multivariate regression technique)</p>	<p>Paired ttest (2 groups or time-points)</p> <p>Repeated-measures ANOVA (2 or more groups or time-points)</p> <p>Mixed models/GEE modeling: (multivariate regression techniques)</p>	<p><u>Non-parametric statistics</u></p> <p>Wilcoxon sign-rank test (alternative to the paired ttest)</p> <p>Wilcoxon rank-sum test (alternative to the ttest)</p> <p>Kruskal-Wallis test (alternative to ANOVA)</p> <p>Spearman rank correlation coefficient (alternative to Pearson's correlation coefficient)</p>

Recall: randomized trial of DHA and eczema (within group test)...



Reproduced with permission from: Figure 3 of: Koch C, Dölle S, Metzger M, Rasche C, Jungclas H, Rühl R, Renz H, Worm M. Docosahexaenoic acid (DHA) supplementation in atopic eczema: a randomized, double-blind, controlled trial. *Br J Dermatol*. 2008 Apr;158(4):786-92.



Wilcoxon sign-rank test

Statistical question: Did patients improve in SCORAD score from baseline to 8 weeks?

- What is the outcome variable? SCORAD
- What type of variable is it? Continuous
- Is it normally distributed? **No** (and small numbers)
- Are the observations correlated? **Yes**, it's the same people before and after
- How many time points are being compared? two
- → Wilcoxon sign-rank test

<http://www.socscistatistics.com/tests/signedranks/Default2.aspx>

Wilcoxon sign-rank test mechanics... (within group test)

- 1. Calculate the change in SCORAD score for each participant.
- 2. Rank the absolute values of the changes in SCORAD score from smallest to largest.
- 3. Add up the ranks from the people who improved and, separately, the ranks from the people who got worse.
- 4. The Wilcoxon sign-rank compares these values to determine whether improvements significantly exceed declines (or vice versa).

Example (optional)

			$x_{2,i} - x_{1,i}$	
i	$x_{2,i}$	$x_{1,i}$	sgn	abs
1	125	110	1	15
2	115	122	-1	7
3	130	125	1	5
4	140	120	1	20
5	140	140		0
6	115	124	-1	9
7	140	123	1	17
8	125	137	-1	12
9	140	135	1	5
10	135	145	-1	10

order by absolute difference

			$x_{2,i} - x_{1,i}$			
i	$x_{2,i}$	$x_{1,i}$	sgn	abs	R_i	$\text{sgn} \cdot R_i$
5	140	140		0		
3	130	125	1	5	1.5	1.5
9	140	135	1	5	1.5	1.5
2	115	122	-1	7	3	-3
6	115	124	-1	9	4	-4
10	135	145	-1	10	5	-5
8	125	137	-1	12	6	-6
1	125	110	1	15	7	7
7	140	123	1	17	8	8
4	140	120	1	20	9	9

Exclude

Critical Values of $\pm W$ for Small Samples:

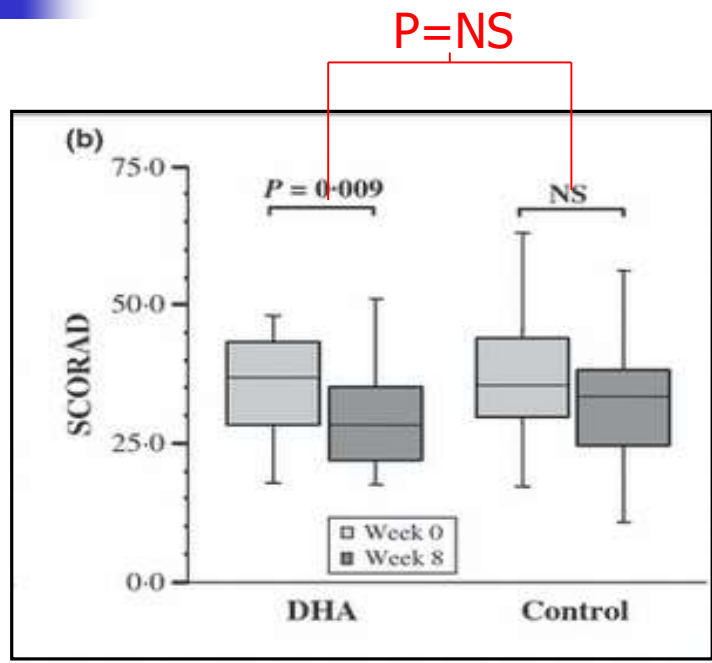
Level of Significance for a				
Directional Test				
	.05	.025	.01	.005
Non-Directional Test				
N	--	.05	.02	.01
5	15	--	--	--
6	17	21	--	--
7	22	24	28	--
8	26	30	34	36
9	29	35	39	43

sgn is the sign function, abs is the absolute value, and R_i is the rank. Notice that pairs 3 and 9 are tied in absolute value. They would be ranked 1 and 2, so each gets the average of those ranks, 1.5.

$N_r = 10 - 1 = 9$, $|W| = |1.5 + 1.5 - 3 - 4 - 5 - 6 + 7 + 8 + 9| = 9$.

$|W| < W_{\alpha=0.05,9,two-sided} = 35 \therefore$ fail to reject H_0 .

Recall: randomized trial of DHA and eczema (between group test)...

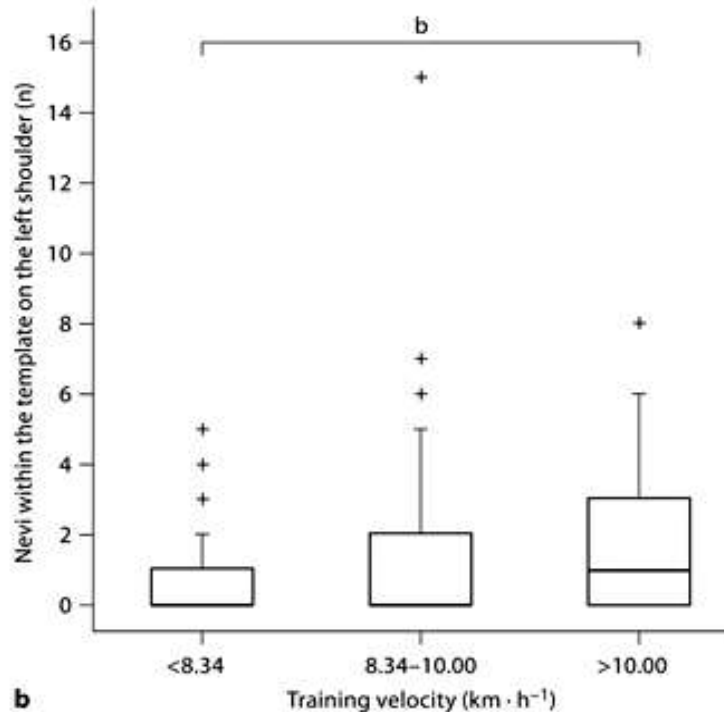


The treatment and placebo groups do not differ, as determined by a Wilcoxon rank-sum test!

Continuous outcome (means)

Outcome Variable	Are the observations independent or correlated?		Alternatives if the normality assumption is violated <u>and</u> small sample size:
	independent	correlated	
Continuous (e.g. pain scale, cognitive function)	Ttest (2 groups) ANOVA (2 or more groups) Pearson's correlation coefficient (1 continuous predictor) Linear regression (multivariate regression technique)	Paired ttest (2 groups or time-points) Repeated-measures ANOVA (2 or more groups or time-points) Mixed models/GEE modeling: (multivariate regression techniques)	<u>Non-parametric statistics</u> Wilcoxon sign-rank test (alternative to the paired ttest) Wilcoxon rank-sum test (alternative to the ttest) Kruskal-Wallis test (alternative to ANOVA) Spearman rank correlation coefficient (alternative to Pearson's correlation coefficient)

Example: Nevi counts and marathon runners



Study: 150 marathon runners volunteered to take part in the skin cancer screening campaign (cross-sectional study). Researchers tested whether the number of nevi (skin lesions that can be precursors to melanoma) is related to sun exposure and training intensity.

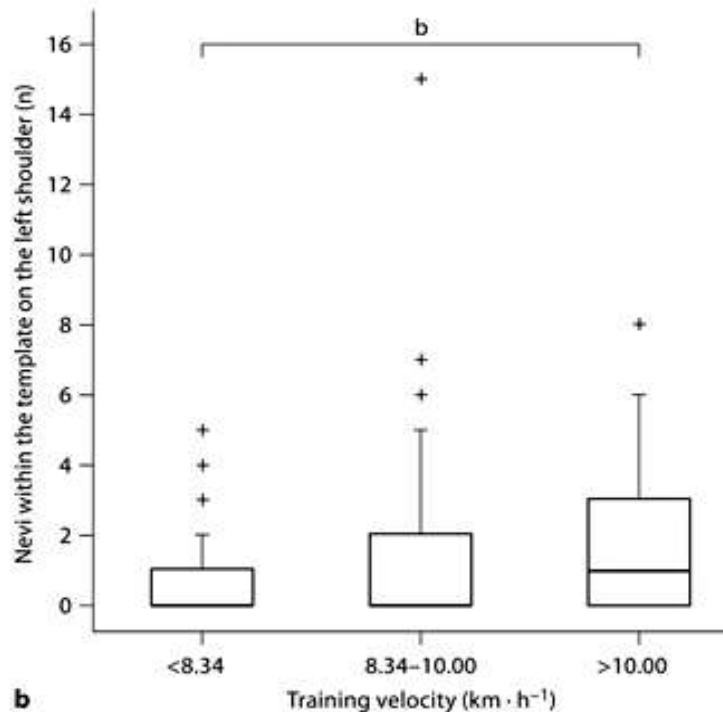
Reproduced with permission from: Richtig et al. Melanoma Markers in Marathon Runners: Increase with Sun Exposure and Physical Strain. *Dermatology* 2008;217:38-44.

Non-parametric ANOVA (Kruskal-Wallis test)

Statistical question: Do nevi counts differ by training velocity (slow, medium, fast) group in marathon runners?

- What is the outcome variable? Nevi count
- What type of variable is it? Continuous
- Is it normally distributed? **No** (and small sample size)
- Are the observations correlated? No
- Are groups being compared and, if so, how many? Yes, three
- → non-parametric ANOVA

Example: Nevi counts and marathon runners



By non-parametric ANOVA, the groups differ significantly in nevi count ($p < .05$) overall.

By Wilcoxon rank-sum test (adjusted for multiple comparisons), the lowest velocity group differs significantly from the highest velocity group ($p < .05$)

Reproduced with permission from: Richtig et al. Melanoma Markers in Marathon Runners: Increase with Sun Exposure and Physical Strain. *Dermatology* 2008;217:38-44.



Statistics in Medicine

Module 4:

Comparing proportions between 2
groups (2x2 table)

Binary or categorical outcomes (proportions)

Outcome Variable	Are the observations correlated?		Alternatives if sparse data:
	independent	correlated	
Binary or categorical (e.g. fracture, yes/no)	Risk difference/relative risks (2x2 table) Chi-square test (RxC table) Logistic regression (multivariate regression technique)	McNemar's chi-square test (2x2 table) Conditional logistic regression (multivariate regression technique) GEE modeling (multivariate regression technique)	McNemar's exact test (alternative to McNemar's chi-square, for sparse data) Fisher's exact test (alternative to the chi-square, for sparse data)



Risk difference/relative risks

From an randomized controlled trial of probiotic supplementation during pregnancy to prevent eczema in the infant:

Table 3. Cumulative incidence of eczema at 12 months of age

	<u>Probiotics group</u>	<u>Placebo group</u>	<u>p-value</u>
Cumulative incidence at 12 months	12/33 (36.4%)	22/35 (62.9%)	0.029



Corresponding 2x2 table

<u>Eczema</u>	<u>Treatment Group</u>		
	Treatment	Placebo	
+	12 (36.4%)	22 (62.9%)	34
-	21	13	34
	33	35	68



Risk difference/relative risk

Statistical question: Does the risk of eczema differ in the treatment and control groups?

- What is the outcome variable? Eczema in the first year of life (yes/no)
- What type of variable is it? Binary
- Are the observations correlated? No
- 2x2 table or RxC table? 2x2 table
- Do we have sparse data (expected value of a cell < 5)? No

→ Absolute risk difference or relative risk, or both



Difference in proportions/risk mechanics...

- Shape: Z-distribution
- Standard error:

2x2 table

<u>Eczema</u>	<u>Treatment Group</u>		
	<u>Treatment</u>	<u>Placebo</u>	
+	12 (36.4%)	22 (62.9%)	34
-	21	13	34
	33	35	68

$$Z = \frac{26.5\%}{\sqrt{\frac{.5 \cdot .5}{33} + \frac{.5 \cdot .5}{35}}} = 2.18$$

34/68

p1 = p2

2-tailed p-value = .029



Risk ratios and odds ratios

- Risk ratio: $\frac{36.4\%}{62.9\%} = 0.58$ 95% CI: .34 - .97

- Corresponding odds ratio:

$$\frac{36.4\% / (1 - 36.4\%)}{62.9\% / (1 - 62.9\%)} = 0.34 \quad 95\% \text{ CI: } .13 - .91$$

Adjusted odds ratio from logistic regression...



From an randomized controlled trial of probiotic supplementation during pregnancy to prevent eczema in the infant:

Table 3. Cumulative incidence of eczema at 12 months of age

	<u>Probiotics group</u>	<u>Placebo group</u>	<u>p-value</u>	<u>Adjusted OR(95% CI)</u>	<u>p-value</u>
Cumulative incidence at 12 months	12/33 (36.4%)	22/35 (62.9%)	0.029*	0.243(0.075–0.792)	0.019†

†p value was calculated by multivariable logistic regression analysis adjusted for the antibiotics use, total duration of breastfeeding, and delivery by cesarean section.

Binary or categorical outcomes (proportions)

Outcome Variable	Are the observations correlated?		Alternatives if sparse data:
	independent	correlated	
Binary or categorical (e.g. fracture, yes/no)	Risk difference/relative risks (2x2 table)	McNemar's chi-square test (2x2 table)	McNemar's exact test (alternative to McNemar's chi-square, for sparse data)
	Chi-square test (RxC table)	Conditional logistic regression (multivariate regression technique)	Fisher's exact test (alternative to the chi-square, for sparse data)
	Logistic regression (multivariate regression technique)	GEE modeling (multivariate regression technique)	



Recall: sunscreen study...

- Researchers assigned 56 subjects to apply SPF 85 sunscreen to one side of their faces and SPF 50 to the other prior to engaging in 5 hours of outdoor sports during mid-day.
- Sides of the face were randomly assigned; subjects were blinded to SPF strength.
- Outcome: sunburn



Incorrect analysis...

Table I -- Dermatologist grading of sunburn after an average of 5 hours of skiing/snowboarding ($P = .03$; Fisher's exact test)

Sun protection factor	Sunburned	Not sunburned
85	1	55
50	8	48

The authors use Fisher's exact test to compare 1/56 versus 8/56. But this counts individuals twice and ignores the correlations in the data!



McNemar's test

Statistical question: Is SPF 85 more effective than SPF 50 at preventing sunburn?

- What is the outcome variable? Sunburn (yes/no)
 - What type of variable is it? Binary
 - Are the observations correlated? Yes, split-face trial
 - Are groups being compared and, if so, how many? Yes, two groups (SPF 85 and SPF 50)
 - Are the data sparse? Yes!
- McNemar's test exact test (if bigger numbers, would use McNemar's chi-square test)



Correct analysis of data...

Table 1. Correct presentation of the data from: Russak JE et al. *JAAD* 2010; 62: 348-349. ($P = .016$; McNemar's test).

	<u>SPF-50 side</u>	
	Sunburned	Not sunburned
<u>SPF-85 side</u>		
Sunburned	1	0
Not sunburned	7	48

Only the 7 discordant pairs provide useful information for the analysis!



McNemar's exact test...

- There are 7 discordant pairs; under the null hypothesis of no difference between sunscreens, the chance that the sunburn appears on the SPF 85 side is 50%.
- In other words, we have a binomial distribution with $N=7$ and $p=.5$.
- What's the probability of getting $X=0$ from a binomial of $N=7$, $p=.5$?
- Probability = $\binom{7}{0}.5^7.5^0 = .0078$
- Two-sided probability = $\binom{7}{0}.5^7.5^0 = .0078 + \binom{7}{7}.5^0.5^7 = .0078 = .0156$



McNemar's chi-square test

- Basically the same as McNemar's exact test but approximates the binomial distribution with a normal distribution (works well as long as expected value in each cell ≥ 5)



Statistics in Medicine

Module 5:

Comparing proportions between
more than 2 groups (RxC table)

Binary or categorical outcomes (proportions)

Outcome Variable	Are the observations correlated?		Alternatives if sparse data:
	independent	correlated	
Binary or categorical (e.g. fracture, yes/no)	Risk difference/relative risks (2x2 table)	McNemar's chi-square test (2x2 table)	McNemar's exact test (alternative to McNemar's chi-square, for sparse data)
	Chi-square test (RxC table)	Conditional logistic regression (multivariate regression technique)	Fisher's exact test (alternative to the chi-square, for sparse data)
	Logistic regression (multivariate regression technique)	GEE modeling (multivariate regression technique)	



Recall depression and artery blockage study...

- Relationship between atherosclerosis and late-life depression (Tiemeier et al. *Arch Gen Psychiatry*, 2004).
- Methods: Cross-sectional study. Researchers measured the prevalence of coronary artery calcification (atherosclerosis) and the prevalence of depressive symptoms in a large cohort of elderly men and women in Rotterdam (n=1920).

Results:

Table 3. Relationship Between Coronary Calcifications and Depression Expressed as Odds Ratio*

Atherosclerosis Measure	Subthreshold Depressive Symptoms†				Depressive Disorder†		
	Controls,†						
	No.	No.	Odds Ratio	95% CI	No.	Odds Ratio	95% CI
Coronary calcification							
0-100	865	20	1.0	Reference	9	1.0	Reference
101-500	463	13	1.10	0.53-2.30	11	2.45	0.98-6.13
>500	511	12	0.96	0.43-2.16	16	3.89	1.55-9.77

Abbreviation: CI, confidence interval.

*Odds ratios were calculated with logistic regression adjusted for age, sex, total cholesterol level, cognitive score, blood pressure, diabetes mellitus, body mass index, smoking, antidepressant medication, history of stroke, and myocardial infarction. To test statistical significance of the association between coronary calcifications and depressive disorders, we calculated the overall *P* value with a test for trend; *P* = .004.

†See Table 2 for explanation.



Corresponding RxC table

Coronary calcification level	Number without depression	Number with subclinical deprssion	Number with depressive disorder
Low:	865	20	9
Med:	463	13	11
High:	511	12	16



Chi-square test

Statistical question: Does the risk/prevalence of subclinical depression or depressive disorder differ according to coronary calcification group?

- What is the outcome variable? Depression (none, subclinical, disorder)
 - What type of variable is it? Categorical
 - Are the observations correlated? No
 - 2x2 table or RxC table? RxC table
 - Do we have sparse data (expected value of a cell <5)? No
- Chi-square test

Observed Table:

Coronary calcification	No depression	Subclinical	Clinical depressive disorder	
Low	865	20	9	894
Med	463	13	11	487
High	511	12	16	539
	1839	45	36	1920

Expected Table under Null hypothesis:

Coronary calcification	No depression	Sub-clinical depressive symptoms	Clinical depressive disorder
Low			
Med			
High			



Calculating the expected

- Null hypothesis: variables are independent
- Recall that under independence:
$$P(A)*P(B)=P(A\&B)$$
- Therefore, calculate the marginal probability of B and the marginal probability of A. Multiply $P(A)*P(B)*N$ to get the expected cell count.

Observed Table:

Coronary calcification	No depression	Subclinical	Clinical depressive disorder	
Low	865	20	9	894
Med	463	13	11	487
High	511	12	16	539
	1839	45	36	1920

Expected Table under Null hypothesis:

Coronary calcification	No depression	Sub-clinical depressive symptoms	Clinical depressive disorder
Low	$894 \times 1839 / 1920 =$ 856.3	$894 \times 45 / 1920 =$ 21	$894 - (21 + 856.3) =$ 16.7
Med	$487 \times 1839 / 1920 =$ 466.5	$487 \times 45 / 1920 =$ 11.4	$487 - (466.5 + 11.4) =$ 9.1
High	$1839 - (856.3 + 466.5) =$ 516.2	$45 - (21 + 11.4) =$ 12.6	$36 - (16.7 + 9.1) =$ 10.2



Chi-square test:

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

$$\begin{aligned}\chi_4^2 &= \frac{(865 - 856.3)^2}{856.3} + \frac{(20 - 21)^2}{21} + \frac{(9 - 16.7)^2}{16.7} \\ &+ \frac{(463 - 466.5)^2}{466.5} + \frac{(13 - 11.4)^2}{11.4} + \frac{(11 - 9.1)^2}{9.1} + \\ &\frac{(511 - 516.2)^2}{516.2} + \frac{(12 - 12.6)^2}{12.6} + \frac{(16 - 10.2)^2}{10.2} = 7.877 \\ p &= .096\end{aligned}$$

Degrees of freedom	α									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	18.923	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.262	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Probabilities

Chi-square value

Source: Donald B. Owen, *Handbook of Statistics Tables*, The Chi-Square Distribution Table, © 1962 by Addison-Wesley Publishing Company, Inc. Copyright renewal © 1990. Reprinted by permission of Pearson Education, Inc.



Chi-square p-value calculator online:

	Chi square	DF	
m chi ²	<input type="text" value="7.877"/>	<input type="text" value="4"/>	<input type="button" value="Compute P"/>

P Value Results

Chi²=7.877 DF=4

The two-tailed P value equals 0.0962

By conventional criteria, this difference is considered to be not quite statistically significant.

Binary or categorical outcomes (proportions)

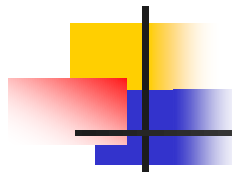
Outcome Variable	Are the observations correlated?		Alternatives if sparse data:
	independent	correlated	
Binary or categorical (e.g. fracture, yes/no)	Risk difference/relative risks (2x2 table)	McNemar's chi-square test (2x2 table)	McNemar's exact test (alternative to McNemar's chi-square, for sparse data)
	Chi-square test (RxC table)	Conditional logistic regression (multivariate regression technique)	Fisher's exact test (alternative to the chi-square, for sparse data)
	Logistic regression (multivariate regression technique)	GEE modeling (multivariate regression technique)	



Fisher's exact test

- In the case of sparse data in an RxC table (expected value of any cells < 5), use the Fisher's exact test.

Fisher's "Tea-tasting experiment"



Claim: Fisher's colleague (call her "Cathy") claimed that, when drinking tea, she could distinguish whether milk or tea was added to the cup first.

To test her claim, Fisher designed an experiment in which she tasted 8 cups of tea (4 cups had milk poured first, 4 had tea poured first).

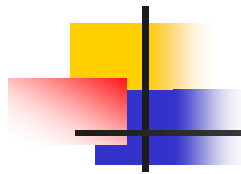
Null hypothesis: Cathy's guessing abilities are no better than chance.

Alternatives hypotheses:

Right-tail: She guesses right more than expected by chance.

Left-tail: She guesses wrong more than expected by chance

Fisher's "Tea-tasting experiment"



Experimental Results:

Guess poured first

Milk

Tea

Poured First

Milk

Tea

3	1
1	3

4

4

Step 1: Identify tables that are as extreme or more extreme than what actually happened:

Here she identified 3 out of 4 of the milk-poured-first teas correctly. The only way she could have done better is if she identified 4 of 4 correct.

<u>Poured First</u>	<u>Guess poured first</u>		
	Milk	Tea	
Milk	3	1	4
Tea	1	3	4

<u>Poured First</u>	<u>Guess poured first</u>			
	Milk	Tea		
Milk	4	0	4	More extreme
Tea	0	4	4	

Step 2: Calculate the probability of the tables (assuming fixed marginals)

<u>Poured First</u>	<u>Guess poured first</u>	
	Milk	Tea
Milk	3	1
Tea	1	3

4
4

$$P(3) = \frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} = .229$$

<u>Poured First</u>	<u>Guess poured first</u>	
	Milk	Tea
Milk	4	0
Tea	0	4

4
4

$$P(4) = \frac{\binom{4}{4}\binom{4}{0}}{\binom{8}{4}} = .014$$

Step 3: to get the left tail and right-tail p-values, consider the probability mass function:

Probability distribution, where X= the number of correct identifications of the milk-poured-first cups:

$$P(4) = \frac{\binom{4}{4}\binom{4}{0}}{\binom{8}{4}} = .014$$

$$P(3) = \frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} = .229$$

$$P(2) = \frac{\binom{4}{2}\binom{4}{2}}{\binom{8}{4}} = .514$$

$$P(1) = \frac{\binom{4}{1}\binom{4}{3}}{\binom{8}{4}} = .229$$

$$P(0) = \frac{\binom{4}{0}\binom{4}{4}}{\binom{8}{4}} = .014$$

“right-hand tail
probability”: $p=.243$

The “two-sided p-value” is calculated by adding up all probabilities in the distribution that are less than or equal to the probability of the observed table (“equal or more extreme”). Here:
 $0.229+.014+.0.229+.014= .4857$

“left-hand tail probability”
(testing the alternative hypothesis that she’s systematically wrong):
 $p=.986$



Statistics in Medicine

Module 6:

Comparing time-to-event outcomes
between 2 or more groups



Recall: time-to-event variables

- The time it takes for an event to occur, if it occurs at all
- Hybrid variable—has a continuous part (time) and a binary part (event: yes/no)
- Only encountered in studies that follow participants over time—such as cohort studies and randomized trials
- Examples:
 - Time to death
 - Time to heart attack
 - Time to chronic kidney disease



Time-to-event variable

- Time part: The time from entry into a study until a subject has a particular outcome or is censored.
- Binary part: Whether or not the subject had the event. Subjects who do not have the event are said to be “censored.” They are counted as event-free for the time they were enrolled in the study.

Time-to-event outcome (survival analysis)

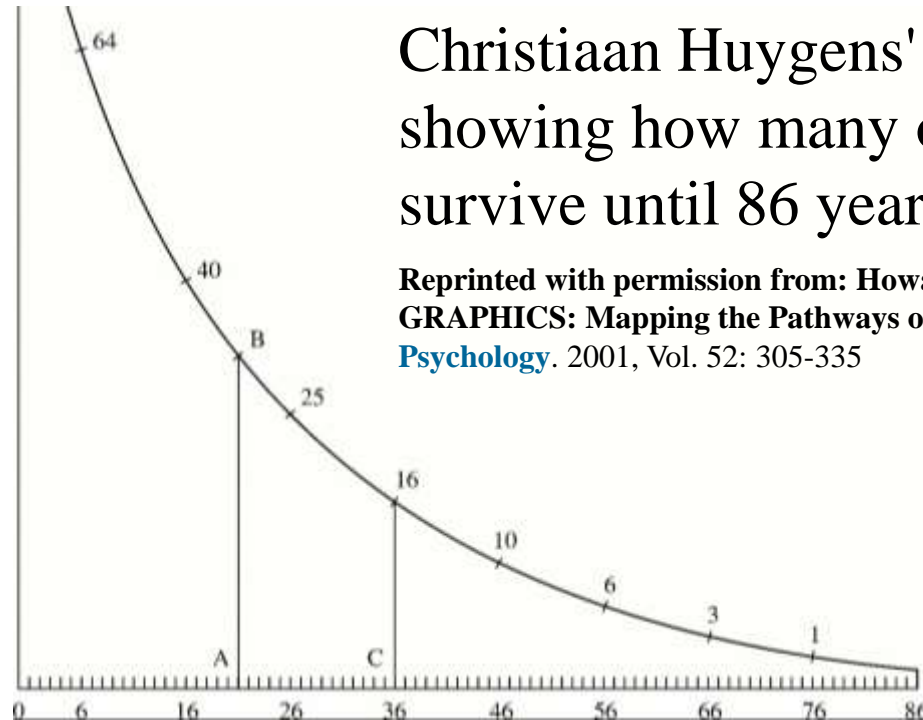
Outcome Variable	Are the observation groups independent or correlated?		Modifications if assumptions violated:
	independent	correlated	
Time-to-event (e.g., time to fracture)	Rate ratio (2 groups) Kaplan-Meier statistics (2 or more groups) Cox regression (multivariate regression technique)	Frailty model (multivariate regression technique)	Time-varying effects



What is survival analysis?

- Statistical methods for analyzing time-to-event data.
- Accommodates data from randomized clinical trial or cohort study design.

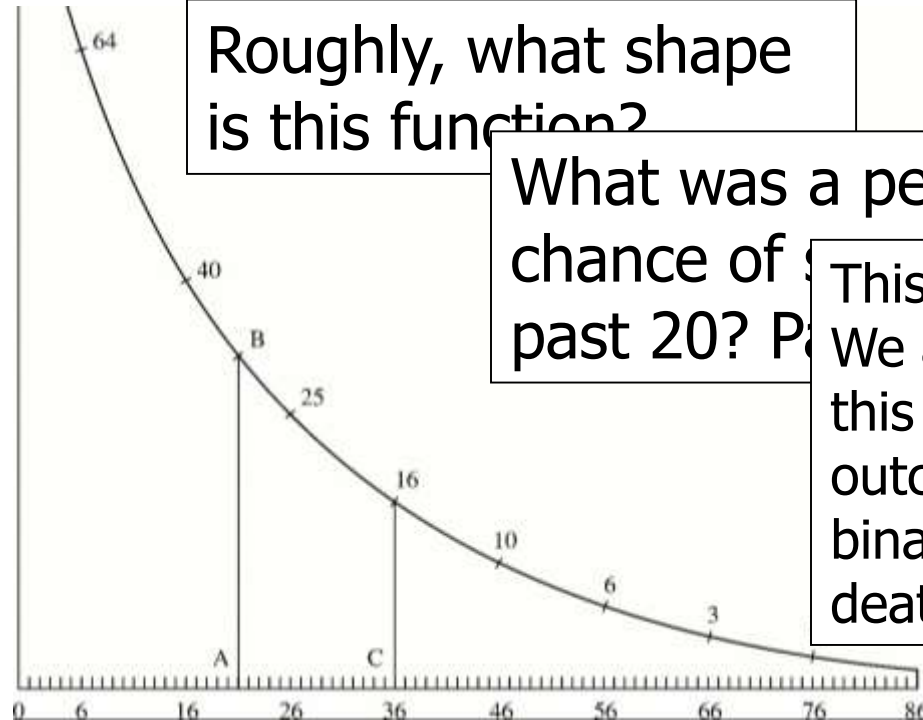
Early example of survival analysis, 1669



Christiaan Huygens' 1669 curve showing how many out of 100 people survive until 86 years.

Reprinted with permission from: Howard Wainer **STATISTICAL GRAPHICS: Mapping the Pathways of Science.** [Annual Review of Psychology](#). 2001, Vol. 52: 305-335

Early example of survival analysis



Roughly, what shape is this function?

What was a person's chance of survival past 20? Past 30? Past 40?

This is survival analysis! We are trying to estimate this curve—only the outcome can be any binary event, not just death.

Time-to-event outcome (survival analysis)

Outcome Variable	Are the observation groups independent or correlated?		Modifications if assumptions violated:
	independent	correlated	
Time-to-event (e.g., time to fracture)	Rate ratio (2 groups) Kaplan-Meier statistics (2 or more groups) Cox regression (multivariate regression technique)	Frailty model (multivariate regression technique)	Time-varying effects



Introduction to Kaplan-Meier

Non-parametric estimate of the survival function:

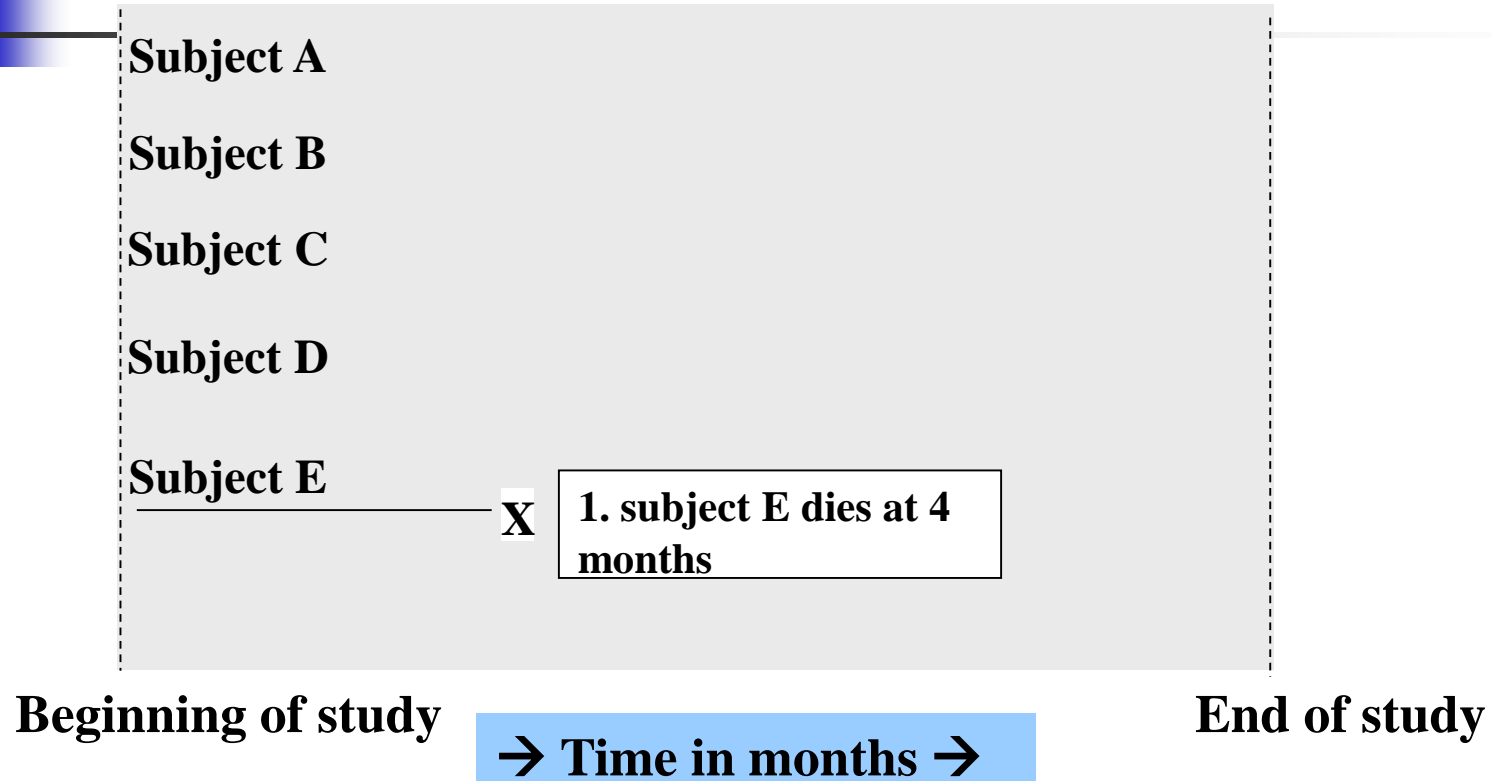
Simply, the empirical probability of surviving past certain times in the sample (taking into account censoring).



Kaplan-Meier Methods

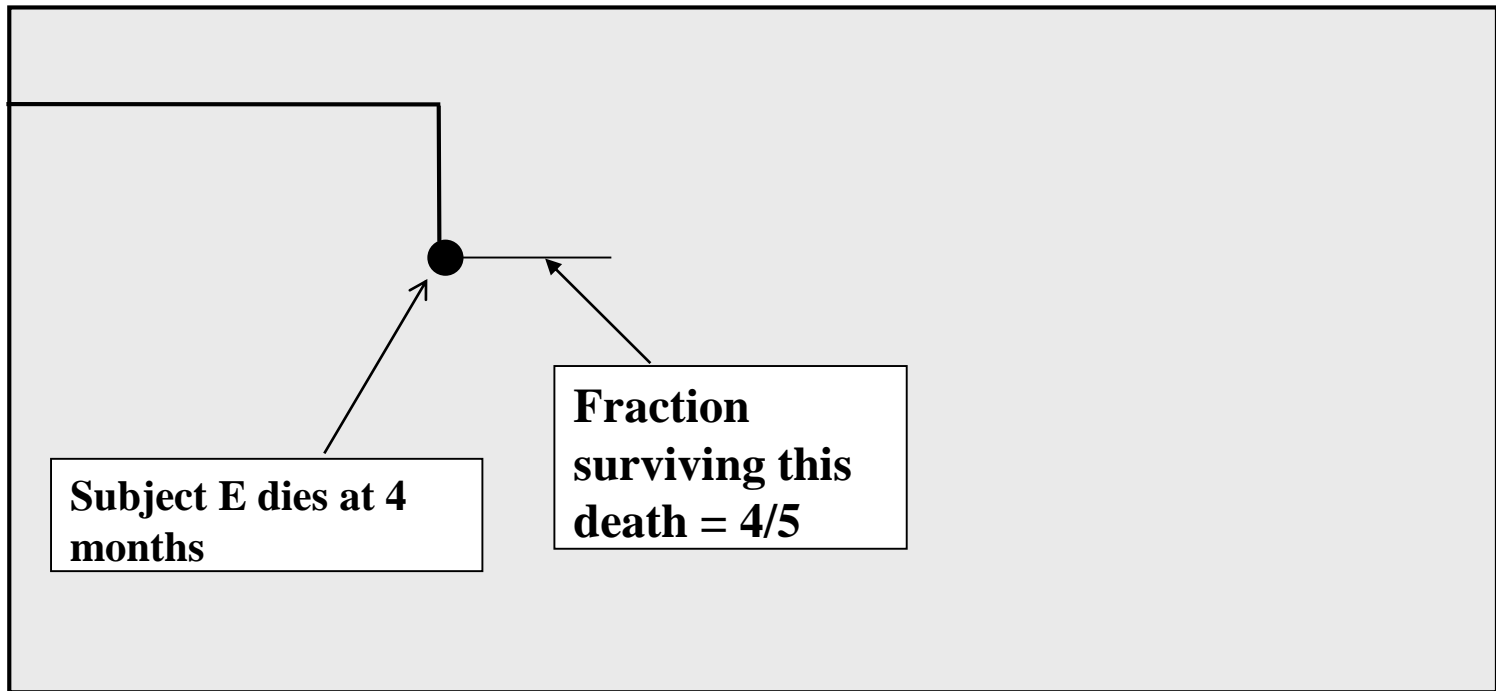
- The empirical probability of surviving past certain times in the sample, taking into account censoring.
- If there was no censoring, the Kaplan-Meier estimate would just be the proportion surviving the study.
- Kaplan-Meier curves for different groups can be statistically compared with a log-rank test (a type of chi-square test).

Hypothetical survival data



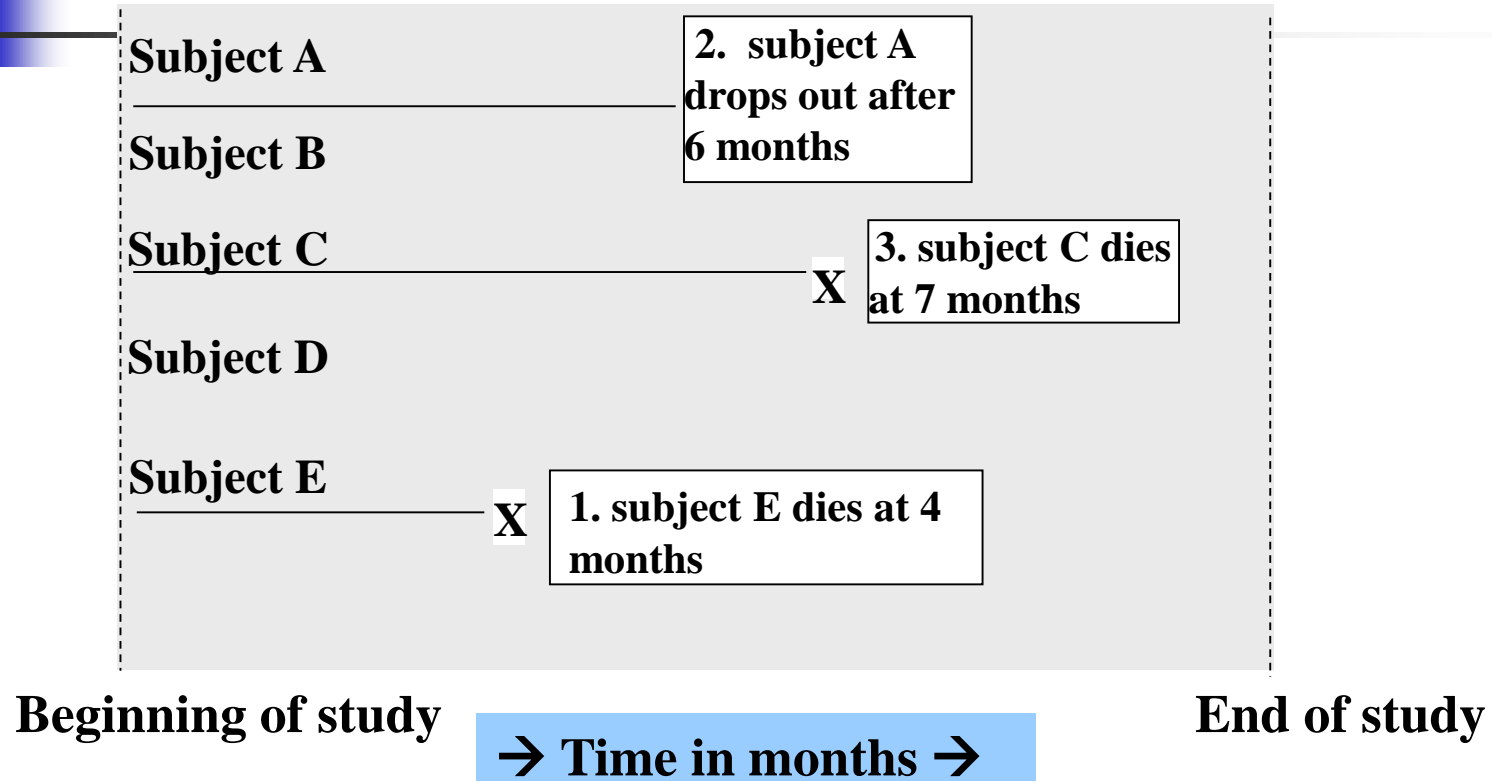
Corresponding Kaplan-Meier Curve

100%



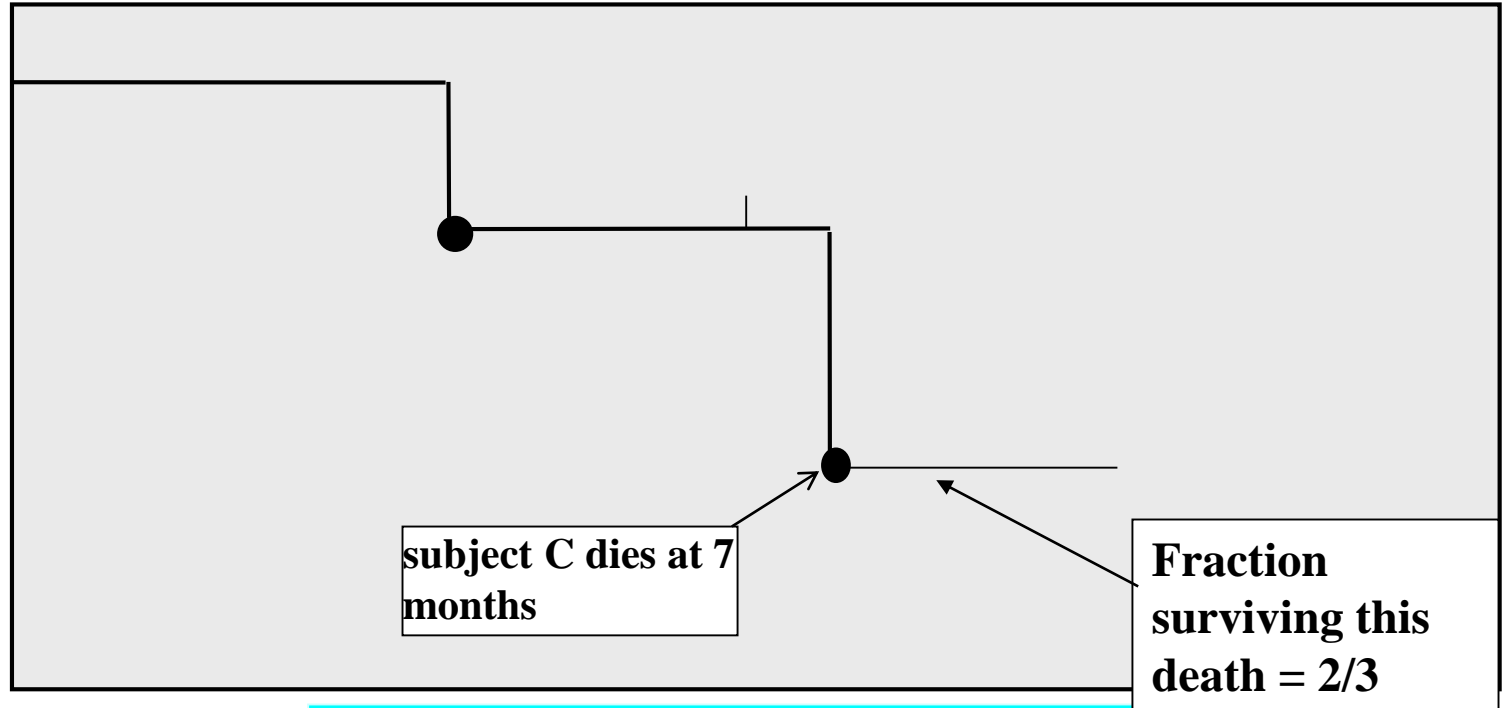
→ Time in months →

Hypothetical survival data



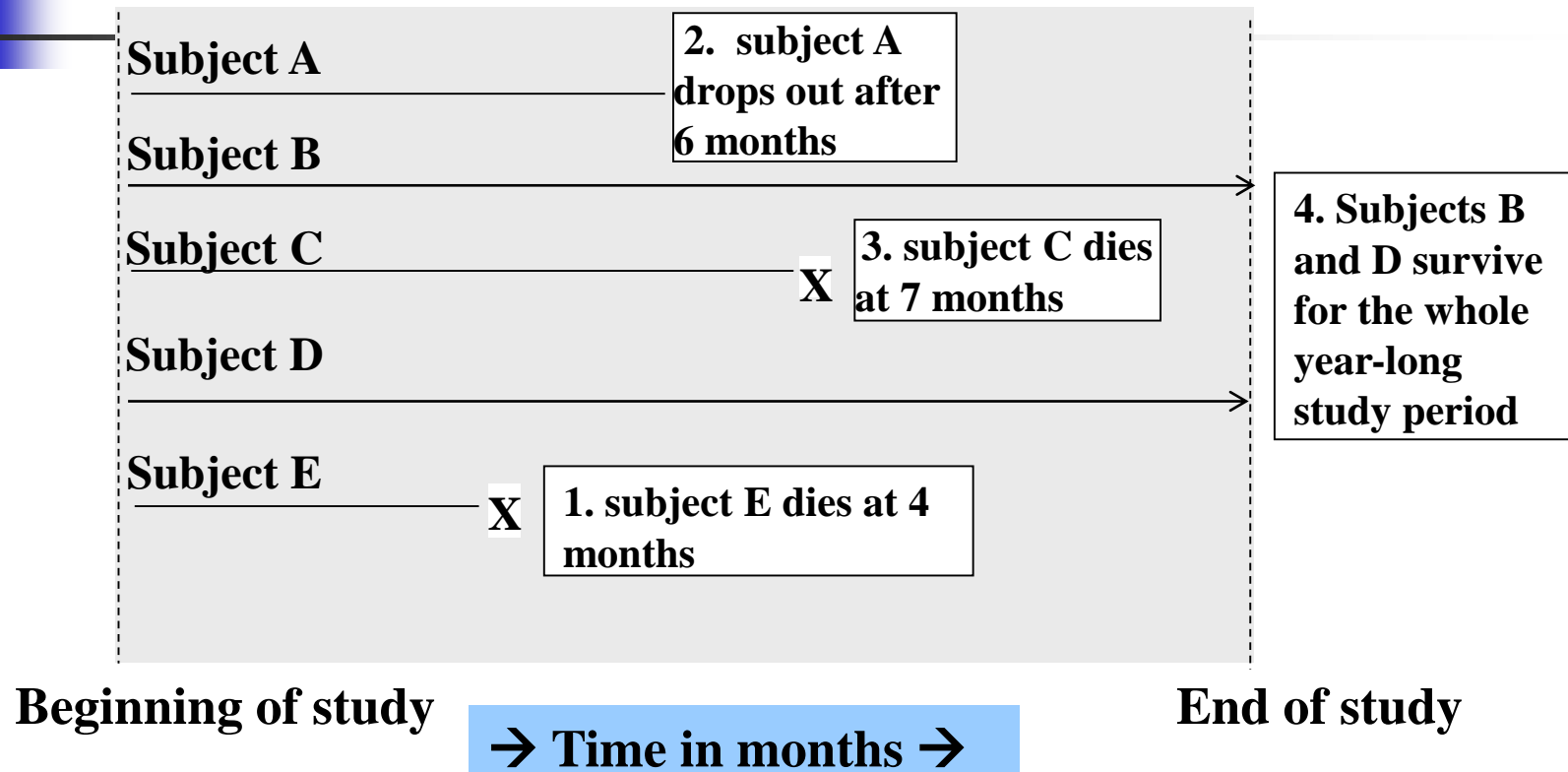
Corresponding Kaplan-Meier Curve

100%

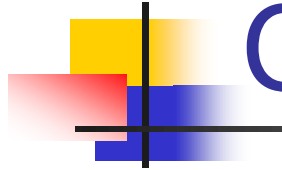


→ Time in months →

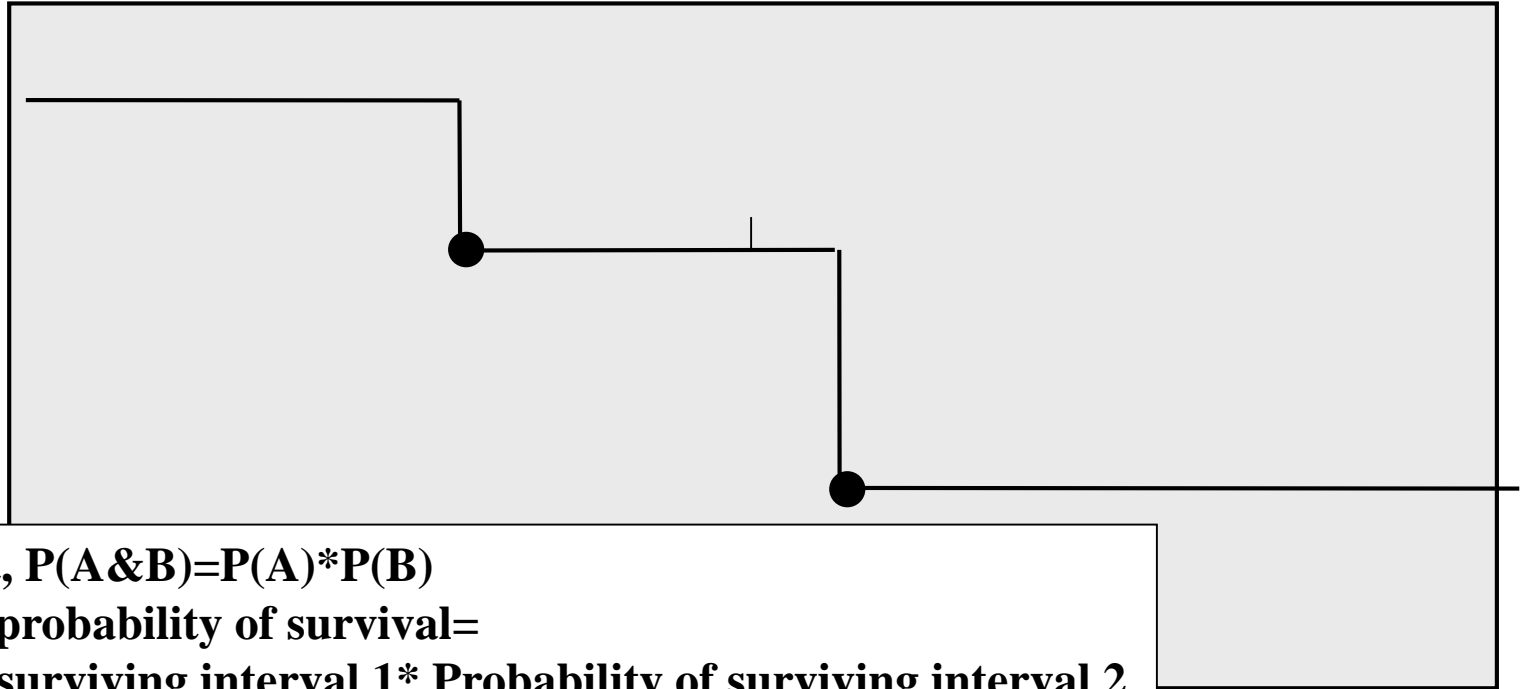
Survival Data



Corresponding Kaplan-Meier Curve



100%



If independent, $P(A \& B) = P(A) * P(B)$

\therefore Cumulative probability of survival =

Probability of surviving interval 1 * Probability of surviving interval 2

$= 4/5 * 2/3 = .5333$



The Kaplan-Meier estimate

- The probability of surviving in the entire year, taking into account censoring
- $= (4/5) (2/3) = 53\%$
- NOTE: $> 40\%$ ($2/5$) because the one drop-out survived at least a portion of the year.
- AND $< 60\%$ ($3/5$) because we don't know if the one drop-out would have survived until the end of the year.



Example: time-to-conception for subfertile women

“Failure” here is a good thing.

38 women (in 1982) were treated for infertility with laparoscopy and hydrotubation.

All women were followed for up to 2-years to describe time-to-conception.

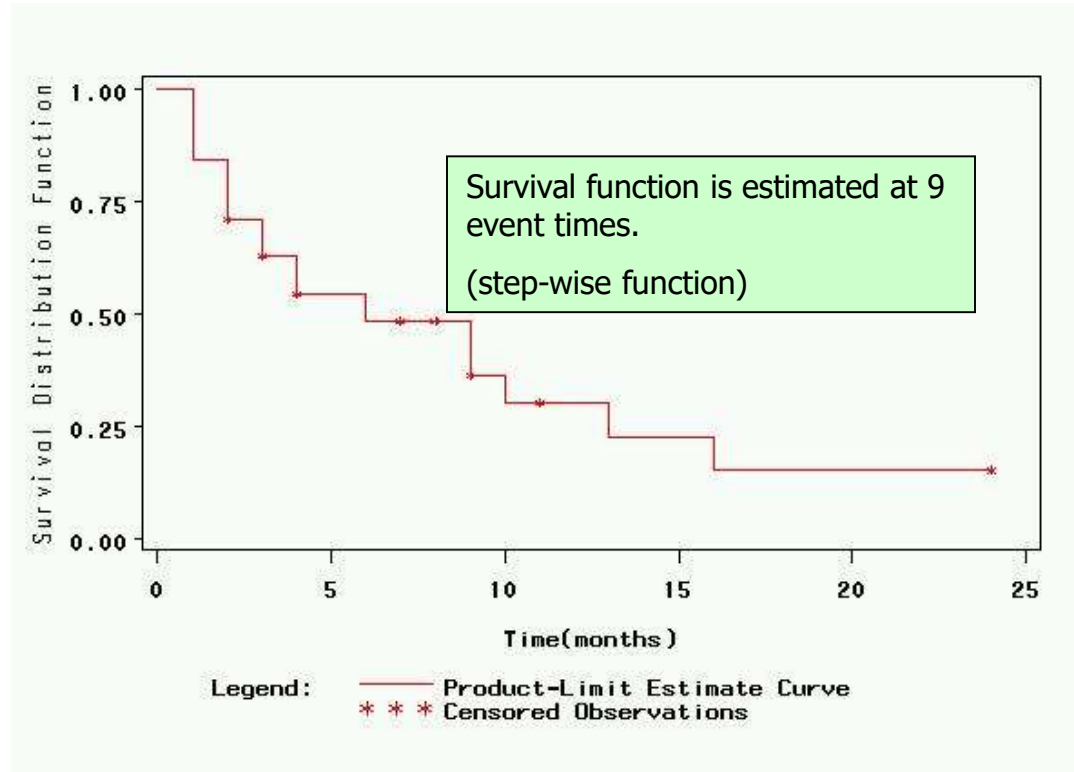
The event is conception, and women "survived" until they conceived.

Data from: BMJ, Dec 1998; 317: 1572 - 1580.

Raw data: Time (months) to conception or censoring in 38 sub-fertile women after laparoscopy and hydrotubation (1982 study):	<u>Conceived (event)</u>	<u>Did not conceive (censored)</u>
	1	2
	1	3
	1	4
	1	7
	1	7
	1	8
	2	8
	2	9
	2	9
	2	9
	2	11
	3	24
	3	24
	3	
	4	
	4	
	4	
	6	
	6	
	9	
	9	
	9	
	10	
	13	
	16	

*Table reproduced with permission
 from: Bland JM, Altman DG. Survival
 probabilities (the Kaplan-Meier
 method). BMJ. 1998;317:1572.*

Corresponding Kaplan-Meier Curve

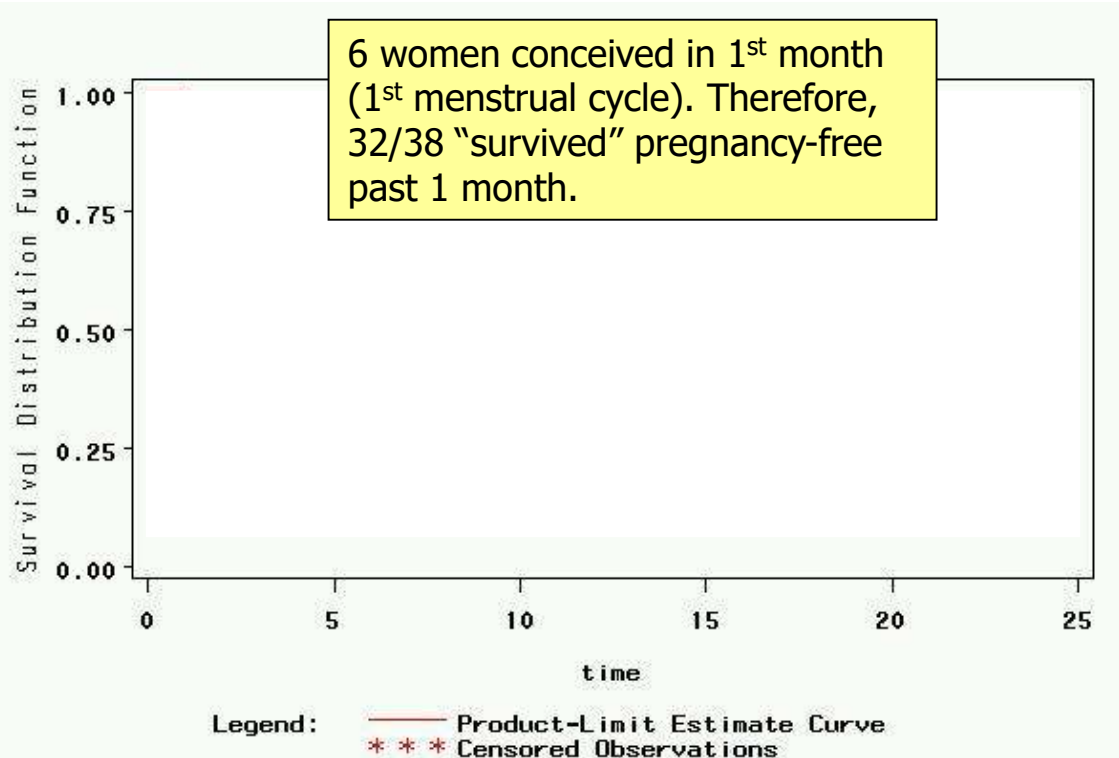


Raw data: Time (months) to
conception or censoring in
38 sub-fertile women after
laparoscopy and
hydrotubation (1982 study):

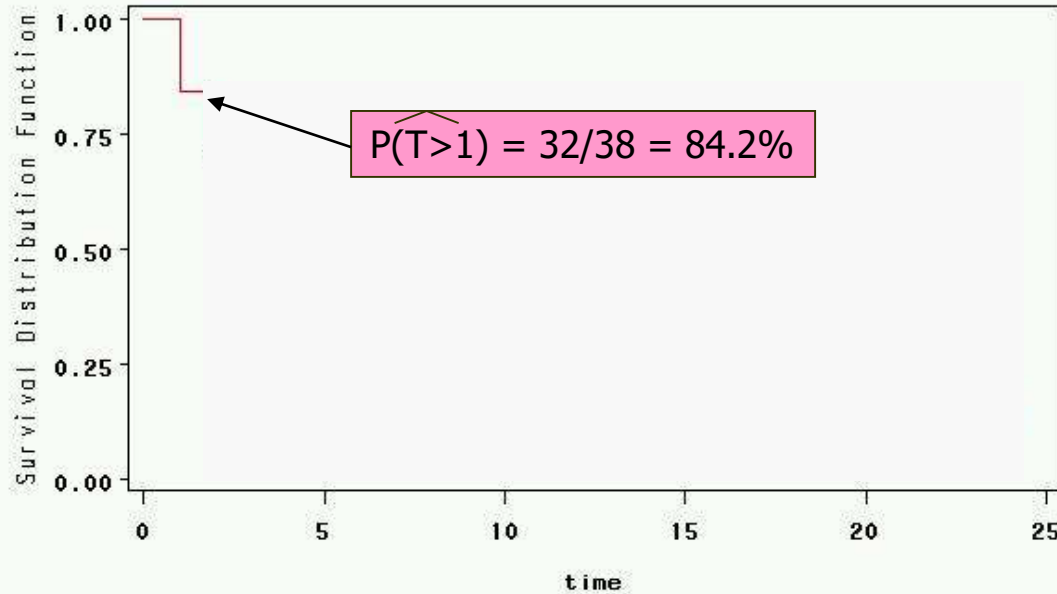
<u>Conceived (event)</u>	<u>Did not conceive (censored)</u>
1	2
1	3
1	4
1	7
1	7
1	8
2	8
2	9
2	9
2	9
2	11
3	24
3	24
3	
4	
4	
4	
4	
6	
6	
9	
9	
9	
10	
13	
16	

Table reproduced with permission
from: Bland JM, Altman DG. Survival
probabilities (the Kaplan-Meier
method). *BMJ*. 1998;317:1572.

Corresponding Kaplan-Meier Curve



Corresponding Kaplan-Meier Curve



Legend: — Product-Limit Estimate Curve
* * * Censored Observations

Raw data: Time (months) to conception or censoring in 38 sub-fertile women after laparoscopy and hydrotubation (1982 study):

<u>Conceived (event)</u>	<u>Did not conceive (censored)</u>
1	2.1
1	3
1	4
1	7
1	7
1	8
2	8
2	9
2	9
2	9
2	11
3	24
3	24
3	
4	
4	
4	
6	
6	
9	
9	
9	
10	
13	
16	

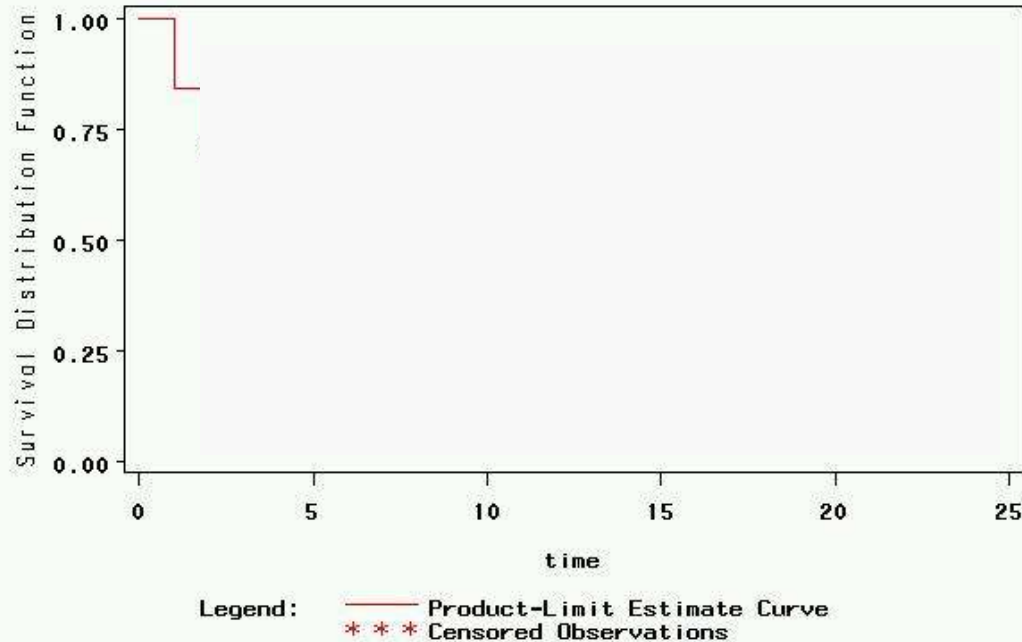
Table reproduced with permission from: Bland JM, Altman DG. Survival probabilities (the Kaplan-Meier method). *BMJ*. 1998;317:1572.

Important detail of how the data were coded:
Censoring at t=2 indicates survival PAST the 2nd cycle (i.e., we know the woman “survived” her 2nd cycle pregnancy-free).

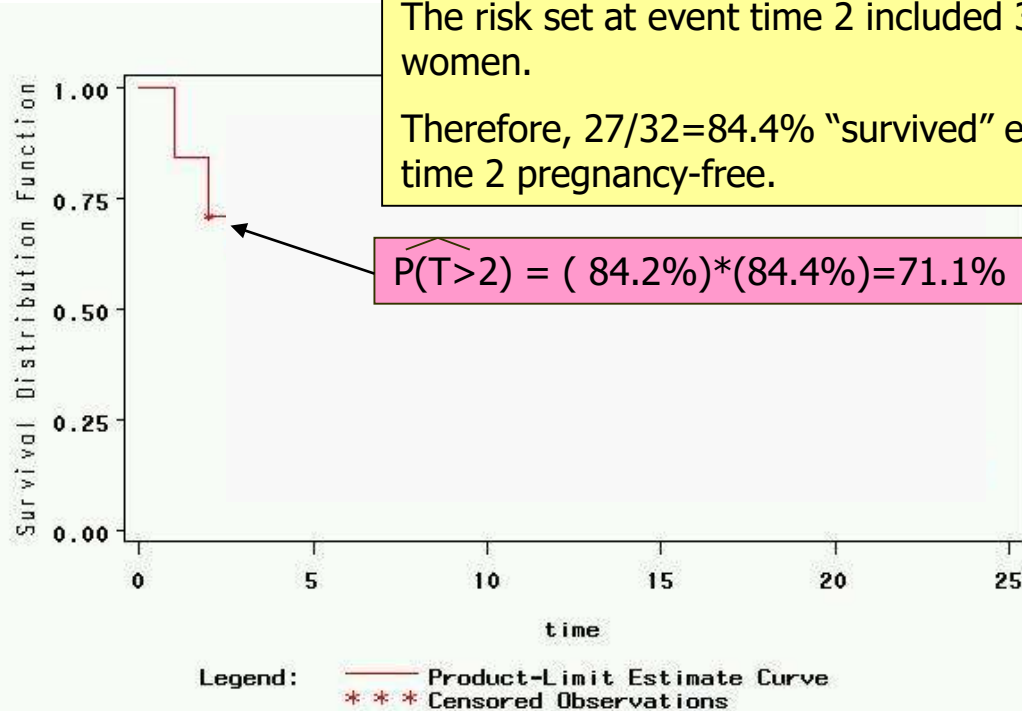
Thus, for calculating KM estimator at 2 months, this person should still be included in the risk set.

Think of it as 2+ months, e.g., 2.1 months.

Corresponding Kaplan-Meier Curve



Corresponding Kaplan-Meier Curve



Raw data: Time (months) to conception or censoring in 38 sub-fertile women after laparoscopy and hydrotubation (1982 study):

Table reproduced with permission from: Bland JM, Altman DG. Survival probabilities (the Kaplan-Meier method). *BMJ*. 1998;317:1572.

Conceived (event)

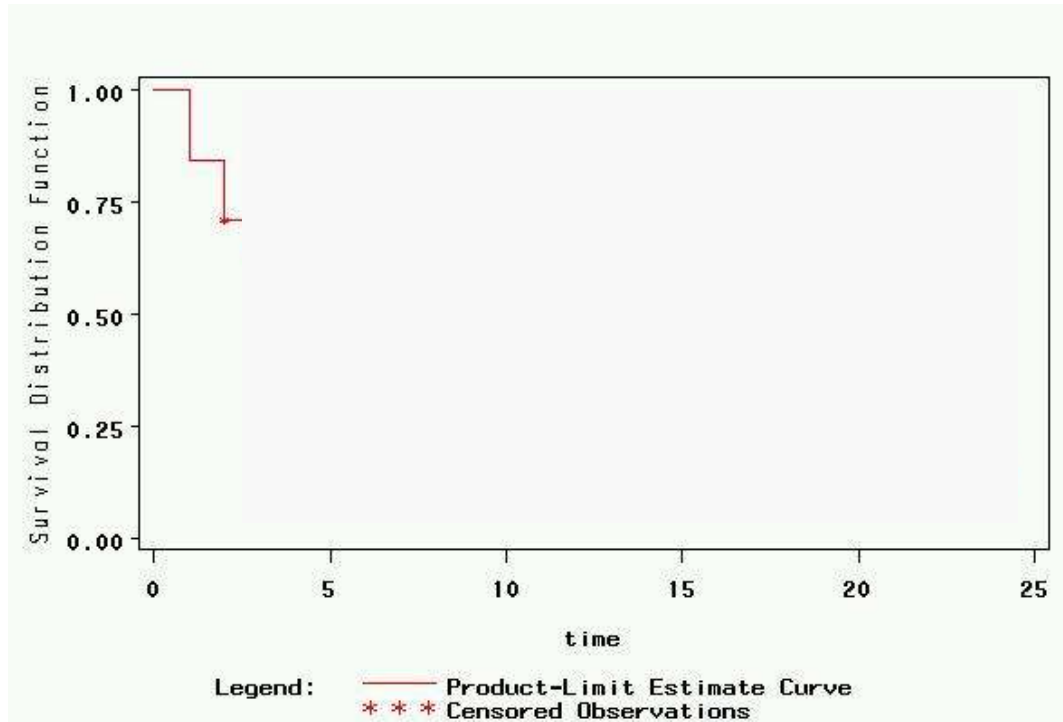
- 1
- 1
- 1
- 1
- 1
- 1
- 2
- 2
- 2
- 2
- 2
- 2
- 3
- 3
- 3
- 4
- 4
- 4
- 6
- 6
- 9
- 9
- 9
- 10
- 13
- 16

Did not conceive (censored)

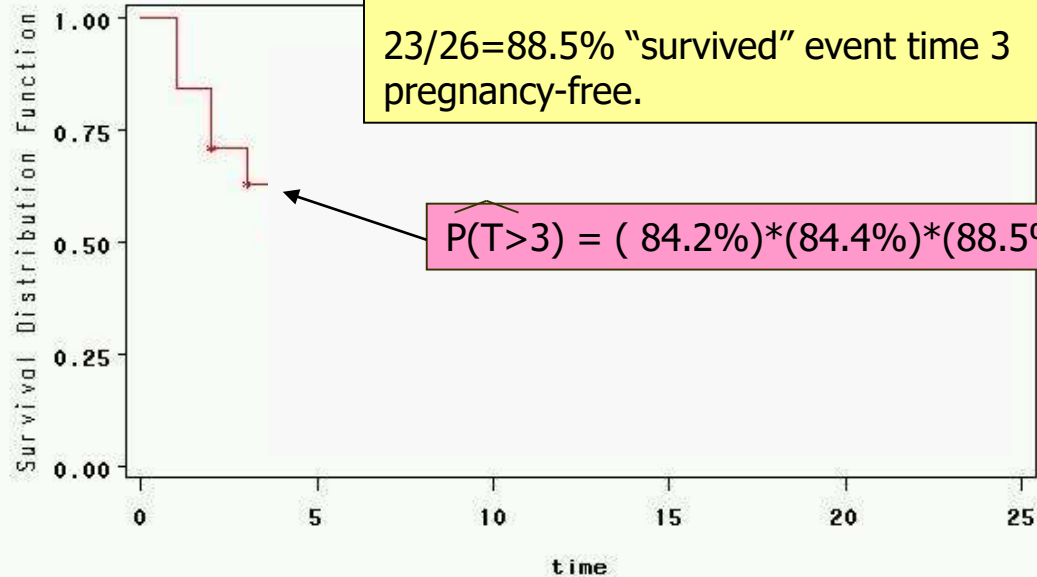
- 2
- 3.1
- 4
- 7
- 7
- 8
- 8
- 9
- 9
- 9
- 11
- 24
- 24

Risk set at 3 months includes 26 women

Corresponding Kaplan-Meier Curve



Corresponding Kaplan-Meier Curve



3 women conceive in the 3rd month.

The risk set at event time 3 included 26 women.

23/26=88.5% "survived" event time 3 pregnancy-free.

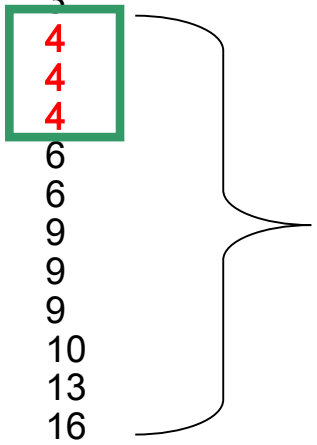
$$P(T>3) = (84.2\%)*(84.4\%)*(88.5\%)=62.8\%$$

Legend: — Product-Limit Estimate Curve
* * * Censored Observations

Raw data: Time (months) to conception or censoring in 38 sub-fertile women after laparoscopy and hydrotubation (1982 study):

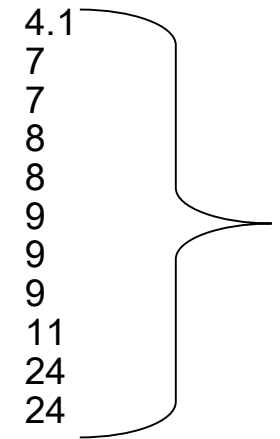
Conceived (event)

- 1
- 1
- 1
- 1
- 1
- 1
- 2
- 2
- 2
- 2
- 2
- 2
- 3
- 3
- 3
- 4
- 4
- 4
- 6
- 6
- 9
- 9
- 9
- 10
- 13
- 16



Did not conceive (censored)

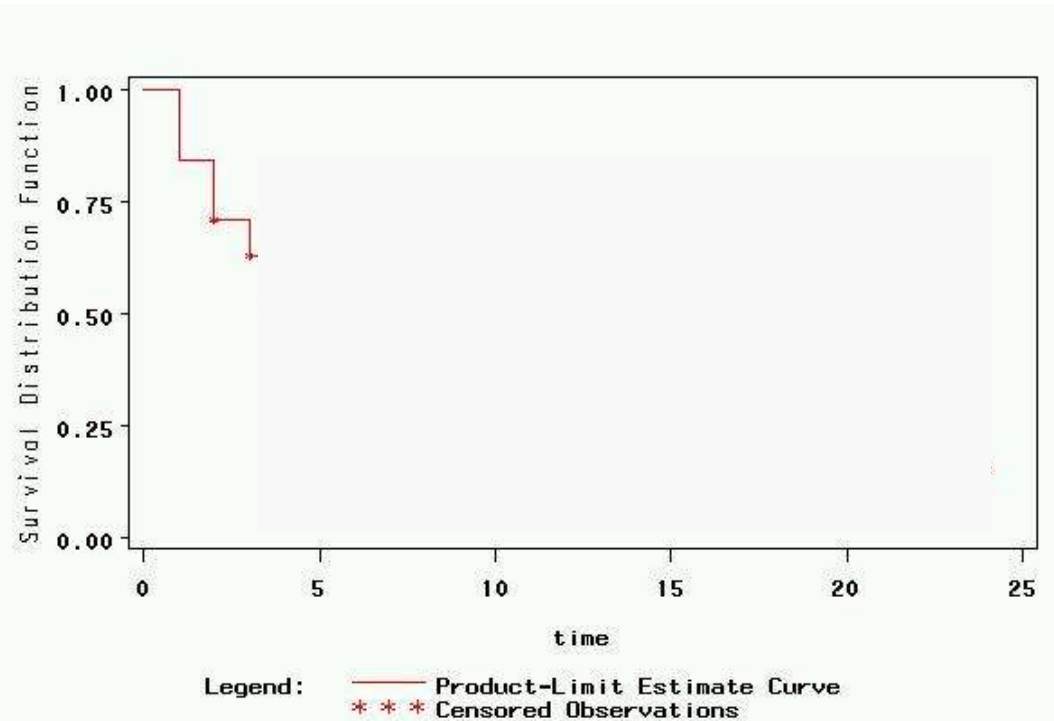
- 2
- 3
- 4.1
- 7
- 7
- 8
- 8
- 9
- 9
- 9
- 11
- 24
- 24



Risk set at 4 months includes 22 women

Table reproduced with permission from: Bland JM, Altman DG. Survival probabilities (the Kaplan-Meier method). *BMJ*. 1998;317:1572.

Corresponding Kaplan-Meier Curve



Corresponding Kaplan-Meier Curve

Survival Distribution Function

1.00
0.75
0.50
0.25
0.00

0

5

10

15

20

25

time

Legend:

— Product-Limit Estimate Curve
* * * Censored Observations

3 women conceive in the 4th month, and 1 was censored between months 3 and 4.

The risk set at event time 4 included 22 women.

19/22=86.4% "survived" event time 4 pregnancy-free.

$$\widehat{P}(T > 4) = (84.2\%) * (84.4\%) * (88.5\%) * (86.4\%) = 54.2\%$$

Raw data: Time (months) to conception or censoring in 38 sub-fertile women after laparoscopy and hydrotubation (1982 study):

Conceived (event)

- 1
- 1
- 1
- 1
- 1
- 1
- 2
- 2
- 2
- 2
- 2
- 2
- 3
- 3
- 3
- 4
- 4
- 4
- 6
- 6
- 9
- 9
- 9
- 10
- 13
- 16

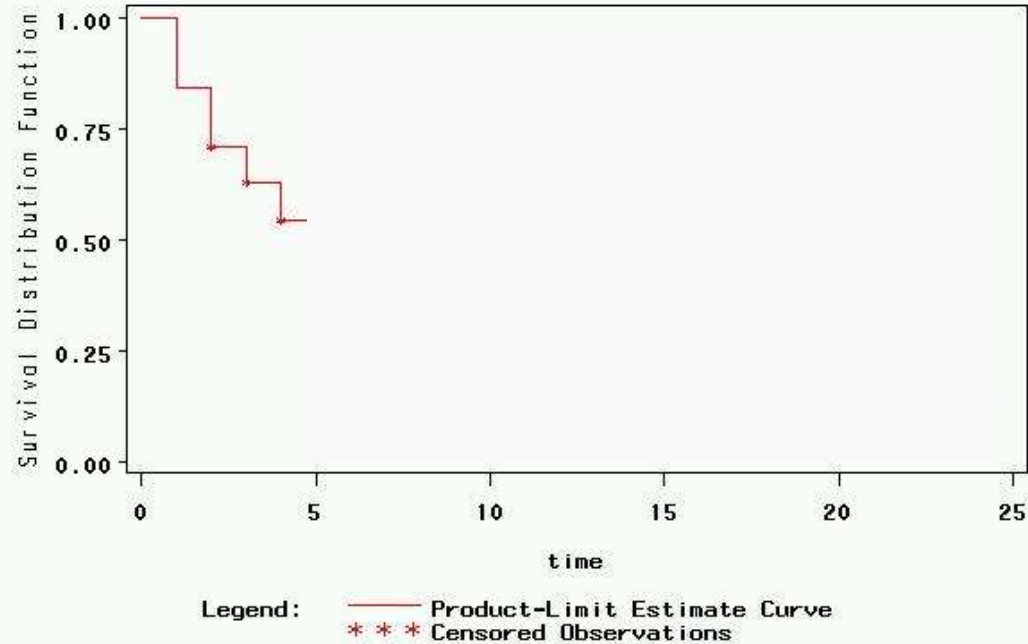
Did not conceive (censored)

- 2
- 3
- 4
- 7
- 7
- 8
- 8
- 9
- 9
- 9
- 11
- 24
- 24

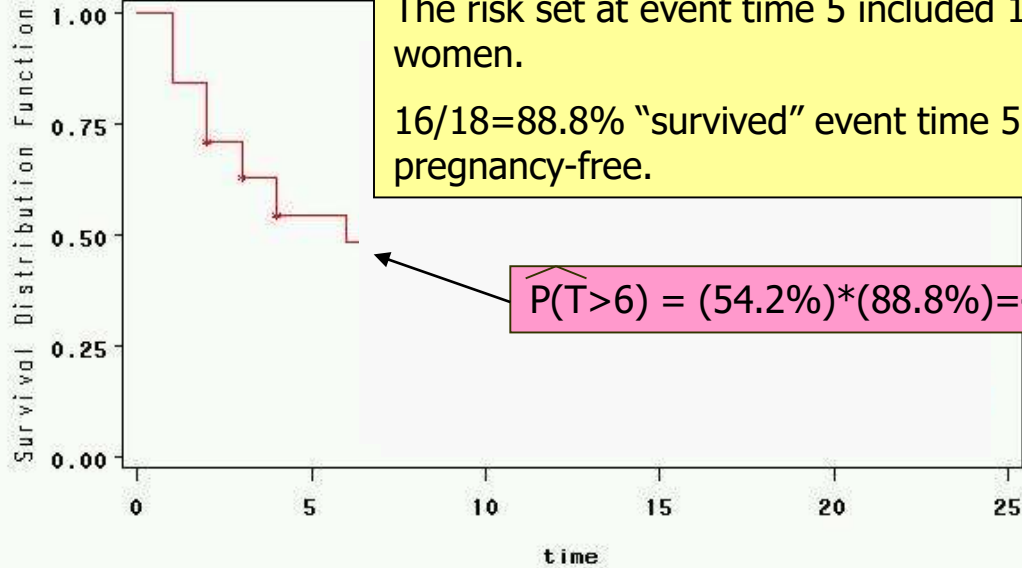
Table reproduced with permission from: Bland JM, Altman DG. Survival probabilities (the Kaplan-Meier method). *BMJ*. 1998;317:1572.

Risk set at 6 months includes 18 women

Corresponding Kaplan-Meier Curve



Corresponding Kaplan-Meier Curve



2 women conceive in the 6th month of the study, and one was censored between months 4 and 6.

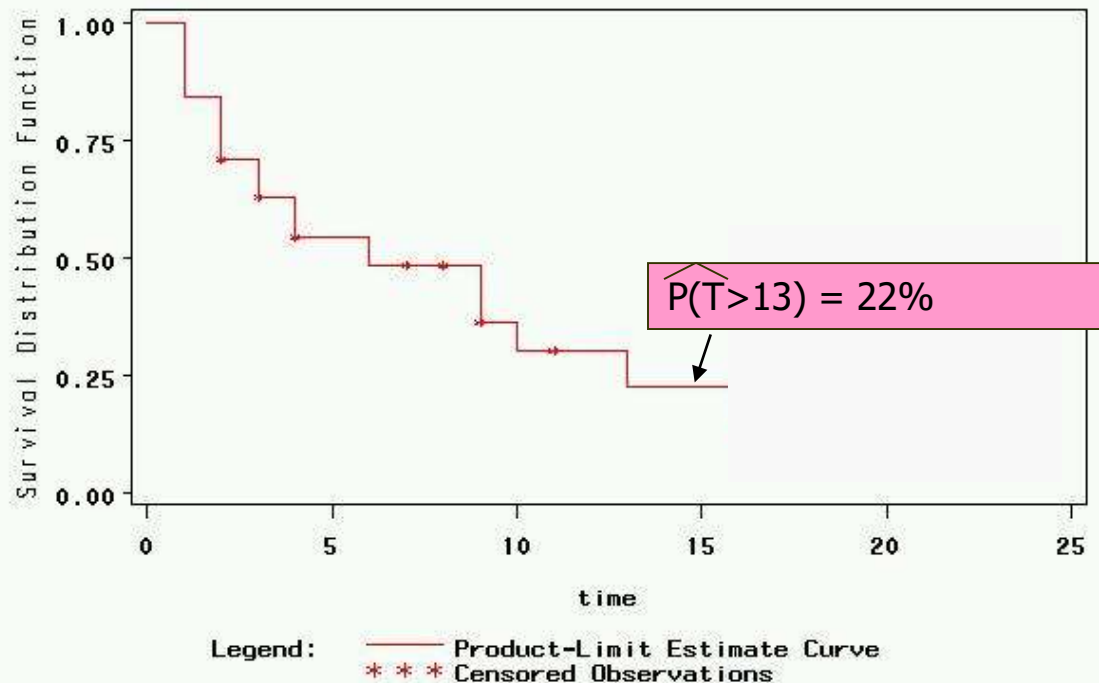
The risk set at event time 5 included 18 women.

16/18=88.8% "survived" event time 5 pregnancy-free.

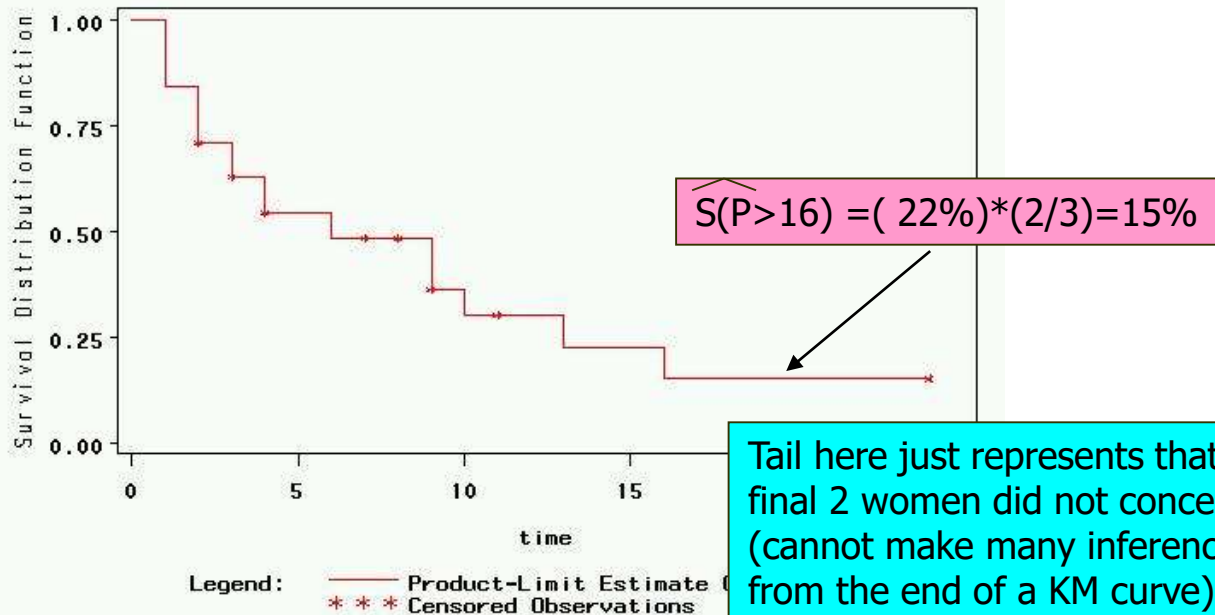
$$\widehat{P}(T>6) = (54.2\%)*(88.8\%)=42.9\%$$

Legend: — Product-Limit Estimate Curve
* * * Censored Observations

Skipping ahead to the 9th and final event time (months=16)...



Skipping ahead to the 9th and final event time (months=16)...



Raw data: Time (months) to conception or censoring in 38 sub-fertile women after laparoscopy and hydrotubation (1982 study):

Table reproduced with permission from: Bland JM, Altman DG. Survival probabilities (the Kaplan-Meier method). *BMJ*. 1998;317:1572.

Conceived (event)

- 1
- 1
- 1
- 1
- 1
- 1
- 2
- 2
- 2
- 2
- 2
- 2
- 3
- 3
- 3
- 3
- 4
- 4
- 4
- 4
- 6
- 6
- 9
- 9
- 9
- 10
- 13
- 16

Did not conceive (censored)

- 2
- 3
- 4
- 7
- 7
- 8
- 8
- 9
- 9
- 9
- 11
- 24
- 24

} 2 remaining at 16 months (9th event time)

Risk set at 15 months includes 3 women



Kaplan-Meier example: comparing 2 groups

Researchers randomized 44 patients with chronic active hepatitis were to receive prednisolone or no treatment (control), then compared survival curves.

Example from: Bland and Altman. Time to event (survival) data. *BMJ* 1998;317:468.

Prednisolone (n=22)

Control (n=22)

2

2

6

3

12

4

54

7

56 *

10

68

22

89

28

96

29

*Indicates censored value.

96

32

125*

37

128*

40

131*

41

140*

54

141*

61

143

63

145*

71

146

127*

148*

140*

162*

146*

168

158*

173*

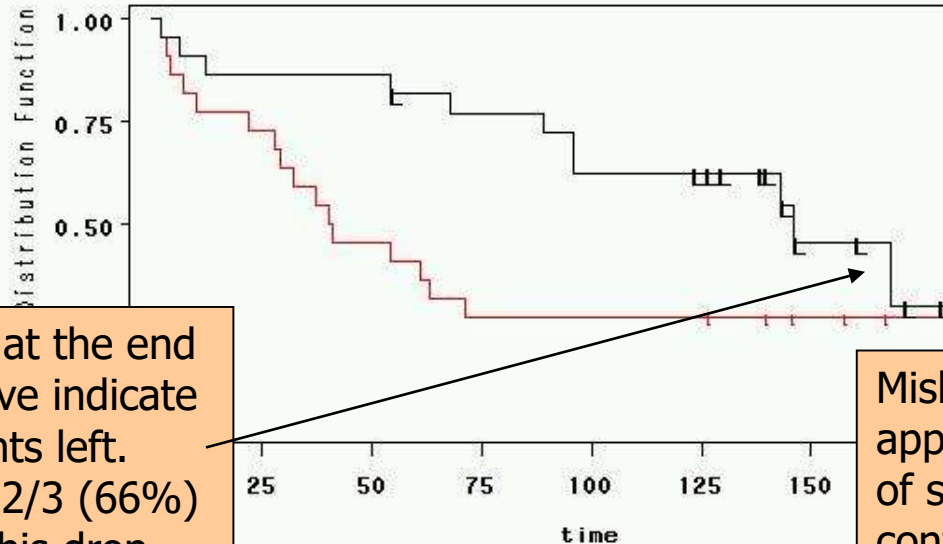
167*

181*

182*

Data reproduced with permission
from Table 1 of: Bland and Altman.
Time to event (survival) data.
BMJ 1998;317:468.

Kaplan-Meier: comparing groups



Are these two curves different?

Big drops at the end of the curve indicate few patients left. E.g., only 2/3 (66%) survived this drop.

Misleading to the eye—apparent convergence by end of study. But this is due to 6 controls who survived fairly long, and 3 events in the treatment group when the sample size was small.

STRATA:

— group=control
L L L Censored group=control
— group=prednisone
L L L Censored group=prednisone



Log-rank test

Test of Equality over Strata

Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	4.6599	1	0.0309

Chi-square test (with 1 degree of freedom) of the (overall) difference between the two groups.

Groups are significantly different.