Biostatistics for Laboratory Animal Veterinarians

Grace E. Kissling, Ph.D.
Biostatistics Branch
National Institute of Environmental Health Sciences
May 18, 2012

My Goals

- To introduce biostatistics concepts and terms
- To provide guidance about when a particular method should or should not be used
- You will not turn into biostatisticians, however ....

Statistical thinking will one day be as necessary for effective citizenship as the ability to read and write
--H. G. Wells

Outline

- Study design
- Levels of measurement
- Numerical and graphical summaries
- Estimation
- Principles of hypothesis testing
- How to choose the right statistical test

Study Design Depends on the Goal

- Comparison of groups
- Comparison with control group
- Trend across dose levels and/or time
- Pre-treatment/Post-treatment change
- Estimation of incidence and/or prevalence
- Associations among several factors
- Prediction of a response from several factors
- Etc., ...

A good study design will..

- Minimize bias
- Minimize confounding
- Maximize statistical power
- Make efficient use of animals, time, other resources

Bias

- The systematic tendency to over-estimate (or under-estimate) a quantity or effect
- Example: A balance has not been calibrated recently and its readings tend to be lower than the actual mass should be.
- Example: An interim sacrifice is planned in an experiment. Only the 'sickest' animals are selected for sacrifice at that time.

Confounding Factor

- An extraneous factor related to both the treatment and the effect that may obscure or exaggerate the true relationship between the treatment and the effect.
- Classical example: There is a significant positive correlation between the crime rate and ice cream sales. Does ice cream consumption cause crime?? Does crime cause people to eat ice cream??

  What is the likely confounder here?

Power

- The probability to detect a difference or effect when it is present.
- We would like the power to be high (>80% or >90%) – more later.....

Design Toolbox

- Randomization
- Blocking/Stratification
- Matching
- Control group
- Factorial layout

Design Tools: Randomization

- Each experimental unit has a chance of being assigned to any treatment.
- Example: Use random number generator to assign animals to groups rather than "haphazardly" reaching in a cage and grabbing the (slowest) animal.
- Reduces bias by distributing pre-existing differences across treatments.
- Without the involvement of a probability process, assumptions of most statistical tests are violated.

Design Tools: Blocking/Stratification

- Experimental units are grouped according to some factor(s). Treatments are assigned at random within each homogeneous group.
- Example: Group animals by body weight before randomization.
- Increases power by separating variation due to an extraneous factor from variation due to the treatment.
- The blocking factor should be related to the response, otherwise you could lose power.

Design Tools: Matching

- A particular form of blocking in which experimental units having common characteristics are paired or matched
- Examples: Twin studies, Pre-post designs
- Just like blocking, matching increases power.

Design Tools: Control Group

- A control group provides a basis for comparison. Usually, it consists of experimental units that do not receive the agent of interest, but they are treated exactly like the other groups in all respects.
- Examples: Untreated controls, Vehicle controls

Design Tools: Factorial Layout

- The effects of two or more factors are studied simultaneously, such that all possible combinations are present in the study.
- Example: Measure the concentration of Compound X in the blood of male and female mice 1 hour after treatment with 0, 50 or 100 mg/kg of X.
- Makes efficient use of resources by allowing study of more than one factor simultaneously.

Design Tools: Factorial Layout

	Compound X (mg/kg)		
	0	50	100
Males			
Females			

Design Tools: Factorial Layout

- Complete factorial: All possible combinations are present.
- Balanced factorial: The sample size is the same* in all cells.
- Incomplete factorial: Some combinations are not included. This may be done on purpose to conserve resources.
  - E.g., Latin square design

* There are exceptions to this definition.

Design Tools: Factorial Layout

- Completely randomized design
  - Experimental units are assigned at random to any one of the cells in the factorial layout.
- Randomized blocks design
  - One factor is a blocking factor (e.g., age group, gender, strain) and experimental units within each level (block) are randomized to treatments.

Putting it all together: Split-plot design

- In the Compound X example, suppose that, because of the animal supplier's constraints, that we have to run the experiment in batches.
  Batch 1 : Oct. 1 delivery, Males only
  Batch 2: Oct. 5 delivery, Females only
  Batch 3: Oct. 7 delivery, Males only
  Etc…
- Blocked by delivery date and gender.
- Randomly assign mice within each batch to a dosage of Compound X.

Putting it all together: Split-plot design

		Compound X (mg/kg)		
Batch (Plots)	Gender	0	50	100
1	Males	✓	✓	✓
2	Females	✓	✓	✓
3	Males	✓	✓	✓
...	...	✓	✓	✓

Further Design Considerations

- What groups/conditions/treatments are to be included?
- When will measurements be taken?
- How many experimental units are needed in each group?

ILAR Journal, Volume 43, No. 4, 2002

How many are needed in each group?

- Depends on several things:
  - Experimental design
  - Effect size, δ
  - Variability, σ
  - Statistical power, 1-β
  - Significance level, α
  - Statistical test to be used
- Formulas, Tables, Charts, Software, Websites are available

Statpages.org is particularly comprehensive

Example

- Does Compound X decrease body weight of mice? How many mice do I need for this experiment?
  - Experimental design: 2 groups, Control and Treated
  - Effect size, δ: a 5 g difference is biologically significant
  - Variability, σ: prior data indicates s.d. = 4 g
  - Statistical power, 1-β: 80%
  - Significance level, α: 0.05, one-sided
  - Statistical test to be used: two-sample t-test

$$n = \frac{2\times(z_\alpha + z_\beta)^2 \times \sigma^2}{\delta^2} = \frac{2\times(1.645+0.84)^2 \times 4^2}{5^2} = 7.9 \rightarrow 8 \text{ per group}$$

Switching gears.....

- We've collected data, now what??
- For any data analysis, the appropriate statistical method will depend on:
  - The design of the study
  - The research question
  - What kind of data are collected

Levels of Measurement

- Nominal
- Ordinal
- Interval/Ratio
- Discrete
- Continuous
- Quantitative
- Qualitative

Levels of Measurement

- Discrete
  - Nominal — yes/no, clinical signs
  - Ordinal grade — 1, 2, 3, 4 or -, +, ++, +++
  - Count — # animals with a liver adenoma
- Continuous
  - Interval/Ratio — mouse body weight; gene expression level

Numerical and Graphical Summaries

- Three important pieces of information:
  - Center/typical value
  - Variability
  - Shape of distribution

Measures of Central Tendency

- Mean – arithmetic average, $\bar{x}$
- Median – half-way point
- Mode – most common value(s)
- Geometric mean – log-based average

$$\bar{x}_{geom} = \sqrt[n]{\prod_{i=1}^{n} x_i} = \exp\left(\frac{\sum \ln(x_i)}{n}\right)$$

Measures of Variability

- Variance
- Standard deviation (SD)
- Standard error of the mean (SE or SEM)
- Coefficient of variation (CV)
- Range
- Interquartile range (IQR)
- Modal percentage
- Geometric standard deviation??
  Tricky! Seek professional help.

Measures of Variability

- Variance
- Standard deviation (SD)
- Standard error of the mean (SE or SEM)
- Coefficient of variation (CV)

$$Var = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1} \qquad SD = \sqrt{Var}$$

$$SEM = \frac{SD}{\sqrt{n}} \qquad CV = \frac{SD}{\bar{x}}$$

Measures of Variability

- Range = Max – Min
- Interquartile range (IQR) = 75th %tile – 25th %tile
- Modal percentage = % in modal category

Standard Deviation or Standard Error?

- SD measures variability of individuals
- SE measures variability of the estimated mean

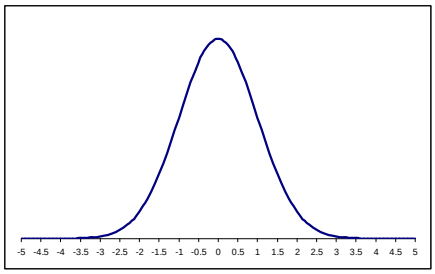
Description Depends on Measurement Level

Level	Center	Variability
Nominal	Mode	Modal percentage
Ordinal	Mode Median	Modal percentage Range, IQR
Interval/ Ratio	Mode Median Mean Geometric mean	Modal percentage Range, IQR SD, SE, CV

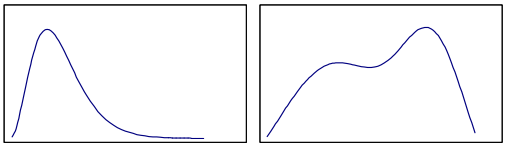
Shape of the Distribution

- Symmetric/Non-symmetric
- Skewed to right/left
- Unimodal/Bimodal/Multimodal
- Normal/Non-normal shape

The Normal/Gaussian/Bell curve Distribution



Non-normal Distributions



Skewed to the right

Bimodal

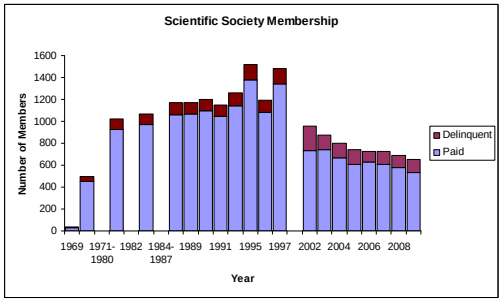
Graphical Methods

A picture is worth a thousand words!

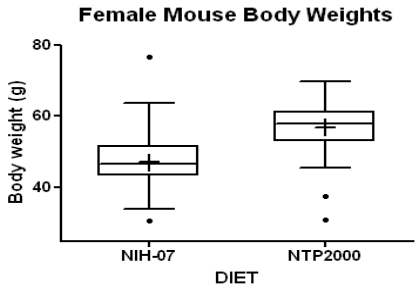
Graphical Methods

- Bar charts
- Histograms
- Boxplots
- Scatter plots
- Q-Q plots

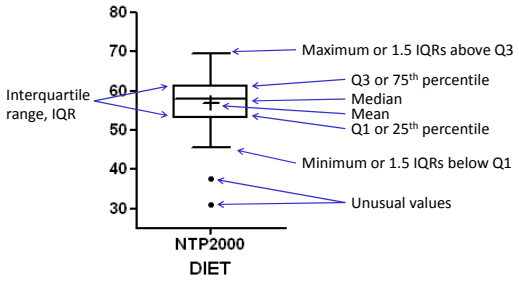
Graphical Methods: Bar Charts



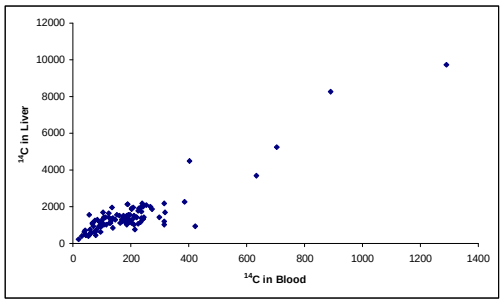
Graphical Methods: Box Plots



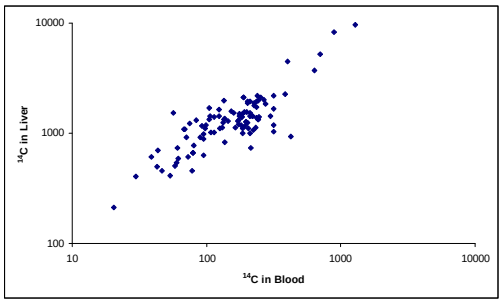
Graphical Methods: Box Plot Components



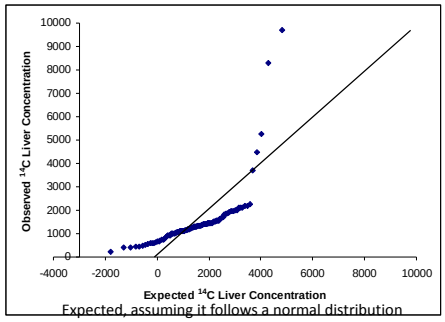
Graphical Methods: Scatter Plots



Graphical Methods: Scatter Plots

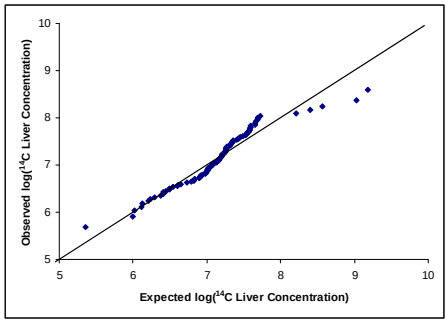


Graphical Methods: Q-Q plots (Quantile-Quantile plots)

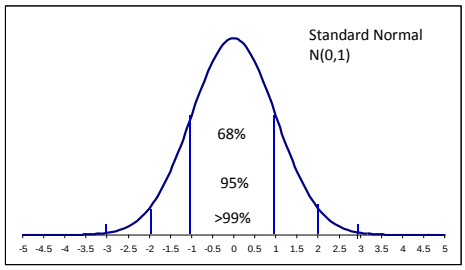


Expected, assuming it follows a normal distribution

Graphical Methods: Q-Q plots (Quantile-Quantile plots)



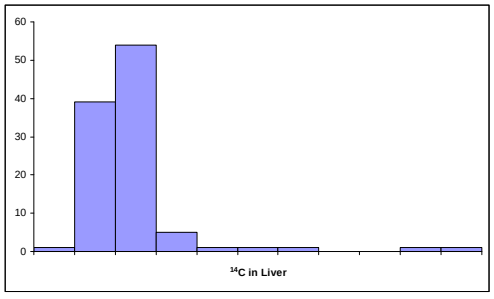
The Normal/Gaussian/Bell curve Distribution



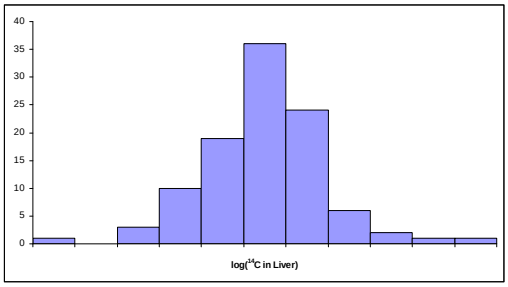
Normality and Transformations

- Many statistical methods rely on having normally distributed data (and sometimes equal variances)
- Normalizing (and variance stabilizing) transformations
  - Logarithmic
  - Square root
  - Box-Cox power transformations, $(x^{\lambda} - 1)/\lambda$

Normality and Transformations: Liver ^{14}C



Normality and Transformations: Log(Liver ^{14}C)

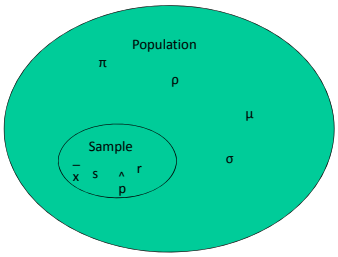


Estimation

A point estimate gives a single value, such as a mean

Interval estimation gives a range of likely values

Population vs. Sample



Confidence Intervals

A 95% confidence interval for the mean gives an interval that has 95% probability of capturing the mean of the population

..... meaning that if we were to conduct the experiment an infinite number of times, 95% of the time, the confidence interval that we construct from the sample will include the mean of the population.

Confidence Intervals

- For the $\log_e(^{14}C$ in liver),
  Mean = 7.14, S.D. = 0.56, N = 104
- The 95% confidence interval for mean $\log(^{14}C$ in liver) is

$$\bar{x} \pm t_{103,.975} \times s.e.m. = 7.14 \pm 1.98 \times 0.56 / \sqrt{104}$$
$$(7.03, 7.25)$$

Exponentiated, the 95% CI for mean ^{14}C in liver is (1130, 1408)

Outliers

- Unusual values that require examination
- ***Do not*** automatically discard outliers
- Several tests exist for detecting outliers
  - Massey-Dixon test or Dixon's Q test
  - Grubbs' test
  - Exceeds 3 standard deviations from the mean
  - Exceeds 1.5 IQRs from 25th or 75th percentiles

Outliers

- Example: ^{14}C in liver (n = 7), Dixon's Q test
- 212, 403, 411, 433, 519, 524, 971

447

759

- Q = 447/759 = 0.59, look up in a Dixon's Q table to see if it exceeds the listed critical values
- For n = 7, the critical values are:
  - 0.507 at p = 0.10
  - 0.568 at p = 0.05 ← 0.59, outlier!
  - 0.680 at p = 0.01

Outliers: Should I remove them?

- Legitimate reasons for removal:
  - Equipment malfunction
  - Impossible value
  - Error in data collection
  - Other mistake in the experiment
- Do not remove if:
  - An unusual value simply can't be explained
  - My data would "look better" if I removed it

Some Summary Measures for Nominal Variables

	Lung cancer	No lung cancer	Total
Smoker	40	80	120
Non-smoker	60	320	380
Total	100	400	500

Relative Risk (RR) = Risk of lung cancer in smokers/
Risk of lung cancer in non-smokers

= (40/120) / (60/380) = 0.33 / 0.16 = 2.09

Some Summary Measures for Nominal Variables

	Lung cancer	No lung cancer	Total
Smoker	40	80	120
Non-smoker	60	320	380
Total	100	400	500

Odds Ratio (OR) = Odds of lung cancer in smokers/
Odds of lung cancer in non-smokers

= (40/120)/(80/120) vs. (60/380)/(320/380)
= 0.50 / 0.19 = 2.63

Screening Tests

Sensitivity = a / (a+c)
Specificity = d / (b+d)

	Disease is Present	Disease is Absent	Total
Test is Positive	a	b	a+b
Test is Negative	c	d	c+d
Total	a+c	b+d	a+b+ c+d

Screening Tests

Sensitivity = a / (a+c)
Specificity = d / (b+d)
Positive Predictive Value = a / (a+b)
Negative Predictive Value = d / (c+d)

	Disease is Present	Disease is Absent	Total
Test is Positive	a	b	a+b
Test is Negative	c	d	c+d
Total	a+c	b+d	a+b+ c+d

Screening Tests

Sensitivity = a / (a+c)
Specificity = d / (b+d)
Positive Predictive Value = a / (a+b)
Negative Predictive Value = d / (c+d)
Efficiency = (a+d)/ (a+b+c+d)

	Disease is Present	Disease is Absent	Total
Test is Positive	a	b	a+b
Test is Negative	c	d	c+d
Total	a+c	b+d	a+b+ c+d

Hypothesis Testing

- Null hypothesis
- Alternative hypothesis
- Test statistic
- P-value
- Conclusion

Hypotheses

- Null hypothesis, H_0
  - No difference, No effect, or No relationship
  - Always test H_0
    - assume H_0 true until there is sufficient evidence to the contrary
    - can not "prove" a hypothesis is true
- Alternative hypothesis, H_1 or H_a
  - Usually, this is the research question

Test Statistic

- This is the evidence in favor of H_0
- Common test statistics have one of 4 well-known distributions:
  - Normal or z
  - Student's t
  - F
  - Chi-square

P-value

- P = Probability of the observed result or results more extreme, assuming H_0 is true
- One-sided or Two-sided P-value?
  - Can you predict *a priori* how groups will differ?
    - YES – use one-sided p
    - NO – use two-sided p

Conclusion

- If p is large, H_0 is supported, we fail to reject H_0
- If p is small, H_0 is not supported, we reject H_0 and conclude that H_a is more likely correct
- We typically use 0.05 to describe what is "large" ($p > 0.05$) and what is "small" ($p < 0.05$).

Keep in Mind....

- Our decision regarding H_0 is based on probabilities, so it could be incorrect
- 0.05 is arbitrary
- We use the significance level and the power to keep the probabilities of an incorrect decision low

	H_0 is TRUE	H_0 is FALSE
Reject H_0	Type I error, False positive α	Correct decision (1 – β), Power
Accept H_0	Correct decision	Type II error, False negative β

Hypothesis Testing: Mouse Body Weights

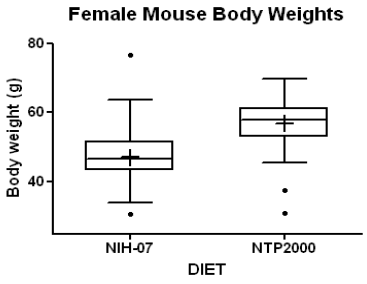
Body weights of 50 female mice on the NIH-07 diet and 50 female mice on the NTP 2000 diet were measured at one year of age.

Is there a difference in mean body weights between the two diet groups?

Two-sided p-value

Hypotheses

- H_0: Mean body weights are the same for the NIH-07 and NTP 2000 diet groups
- H_a: Mean body weights differ between the NIH-07 and NTP 2000 diet groups



Test Statistic and P-value

- NIH-07
  Mean = 47.2g, SD = 7.6g, N = 50
- NTP 2000
  Mean = 56.9g, SD = 7.2g, N = 50

Body weights are typically normally distributed

- Use a two-sample t-test
- $t(98) = 6.55$, $p < 0.0001$ (two-sided)

Conclusion

- If diet has no effect on one-year body weights, the probability of getting a mean difference of 9.7 g or more between the two diets is less than 0.0001.
- Because p is small ($p < 0.0001$), reject H_0 in favor of H_a
- Conclude that there is evidence that mean body weights of female mice at one year differ between the NIH-07 and NTP 2000 diet groups.

Selecting an Appropriate Test

The test statistic selected depends on:

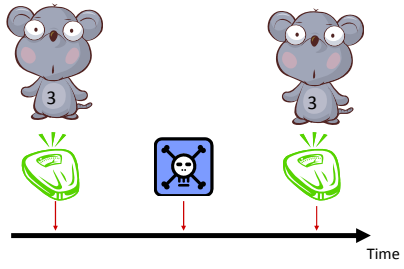
- Study design
- Hypotheses
- Level of measurement of data (nominal, ordinal, interval/ratio)
- Shape of the distribution

A New Study of Body Weights

- H_0: Mean body weights are not affected by treatment with Compound X
- H_a: Mean body weights are decreased after treatment with Compound X

One-sided p-value

Experimental Design



The Data

Mouse	Before X	After X	Difference, After - Before
1	36.3	35.0	-1.3
2	43.5	42.2	-1.3
3	32.0	32.6	0.6
4	50.4	50.6	0.2
5	52.1	51.5	0.6
6	56.3	54.2	-2.1
7	52.4	50.8	-1.6
Mean	46.1	45.3	-0.9
S.D.	9.1	8.7	1.0

Test Statistic and P-value

- Before: Mean = 46.1, SD = 9.1, N = 7
- After: Mean = 45.3, SD = 8.7, N = 7

Body weights are typically normally distributed

- Use a paired t-test
- t(6) = -2.35, p = 0.029 (one-sided)

Conclusion

- Assuming that Compound X has no effect on body weights, the probability of getting an average decrease of 0.9 g or more after administering Compound X is 0.029.
- Because p is small (p = 0.029), reject H_0 in favor of H_a
- Conclude that there is evidence that mean body weights of mice are lower after exposure to Compound X than they were before exposure.

What if I had ignored the study design?

- Before: Mean = 46.1, SD = 9.1, N = 7
- After: Mean = 45.3, SD = 8.7, N = 7

Body weights are typically normally distributed

- Use a two-sample t-test (ignores the pairing)
- t(12) = 0.18, p = 0.429, Not Significant

INCORRECT!!

Hypothesis Testing: Chi-square Test

H_0: Tumor rates are the same in Control and Treated animals
H_a: Tumor rates differ between Control and Treated animals

	Tumor	No Tumor	Total
Control	3	47	50
Treated	10	40	50
Total	13	87	100

χ^2 =4.33 with 1 degree of freedom (df)

P = 0.037

Reject H_0 because 0.037 < 0.05. Conclude that there is a significant difference in tumor rates between Control and Treated animals.

Hypothesis Testing: Fisher's Exact Test

H_0: Tumor rates are the same in Control and Treated animals
H_a: Tumor rates are higher in Treated than Control animals

P = Probability of the observed data or data more extreme, if H_0 is true

	Tumor	No Tumor
Control	3	47
Treated	10	40

	Tumor	No Tumor
Control	2	48
Treated	11	39

	Tumor	No Tumor
Control	1	49
Treated	12	38

	Tumor	No Tumor
Control	0	50
Treated	13	37

Hypothesis Testing: Fisher's Exact Test

	Tumor	No Tumor
Control	3	47
Treated	10	40

Prob = 0.0283

	Tumor	No Tumor
Control	2	48
Treated	11	39

Prob = 0.0064

	Tumor	No Tumor
Control	1	49
Treated	12	38

Prob = 0.0009

	Tumor	No Tumor
Control	0	50
Treated	13	37

Prob = 0.00005

P-value = 0.0283 + 0.0064 + 0.0009 + 0.00005 = 0.0357

Hypothesis Testing: Fisher's Exact Test

p = 0.0357 (one-sided)

Because p = 0.0357 < 0.05, reject H_0 in favor of H_a that the tumor rate is higher in Treated animals than in Control animals.

Hypothesis Testing: Test for Normality

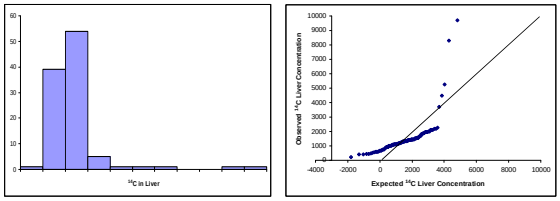
Are my data normally distributed?

H_0: The data are normally distributed.
H_a: The data are not normally distributed.

There are many tests for normality

- Shapiro-Wilks test
- Kolmogorov-Smirnov test
- Liliefors test
- Cramer-von Mises test
- Anderson-Darling test

Test for Normality: Recall ^{14}C in Liver



Hypothesis Testing: Test for Normality

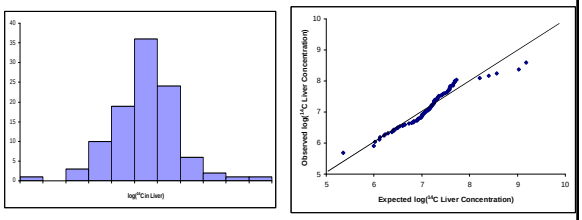
Are my data normally distributed?

H_0: The data are normally distributed.
H_a: The data are not normally distributed.

^{14}C in liver:
Shapiro-Wilks statistic = 0.54
P < 0.0001

Reject H_0: the data are not normally distributed.

Test for Normality: Recall log(^{14}C in Liver)



Hypothesis Testing: Test for Normality

Are my data normally distributed?

H_0: The data are normally distributed.
H_a: The data are not normally distributed.

Log(^{14}C in liver):
Shapiro-Wilks statistic = 0.98
P = 0.44

Accept H_0: the data are normally distributed.

A brief guide to choosing the right statistical test

Nominal		Ordinal		Interval/Ratio				
Independent Groups	Related Groups	Independent Groups	Related Groups	Independent Groups		Related Groups		Assumptions
Tests of proportions		Tests of proportions		Tests of means		Tests of means		
				2 groups	3+ groups	2 groups	3+ groups	
Chi-square test Fisher's exact test	McNemar's chi-square test Kappa statistic	Chi-square test Fisher's exact test	McNemar's chi-square test Kappa statistic Phi coefficient Kendall's tau correlation	Two-sample t-test	Analysis of variance (ANOVA) Dunnett's test (vs. controls) Fisher's LSD test (all pairs)	Paired t-test	Repeated measures ANOVA	Normal distribution
				Mann-Whitney U test Median test	Kruskal-Wallis ANOVA	Wilcoxon signed ranks test Sign test	Friedmans' test	Non-normal distribution
				Tests of variances				
				F ratio of variances	Levene's test Bartlett's test		Sphericity test	Normal distribution
		Tests for trends		Tests for trends				
		Cochran-Armitage trend test		Point-biserial correlation	Linear contrasts from ANOVA		Linear contrasts from repeated measures ANOVA	Normal distribution

		Other tests of association		Other tests of association	
		Phi coefficient		Regression analysis Pearson correlation	Normal distribution
				Poisson regression Binomial regression	Poisson or binomial dist'n
				Robust regression Spearman's rho correlation Kendall's tau correlation	Non-normal distribution
				Tests for normality	
				Kolmogorov-Smirnov test Shapiro-Wilks test Lilliefors test	

Summary

- Every study should be designed to minimize bias and confounding; maximize power; and make the most efficient use of animals, time and resources.
- The appropriate statistical analysis to be used depends on the goal and design of the study, as well as the type of measurements collected.
- An initial, careful descriptive and graphical summary of the data can highlight issues to be considered in the analysis such as outliers, normality, etc.
- Statistical tests of hypotheses rely on probabilities. Hypotheses can not be "proven" with statistics.

Thank you.

Questions?

"IF you torture the data long enough, it will confess. But there is no guarantee that it will tell you the truth."

--from Berk, Regression Analysis: A Constructive Critique, 2004.