A Clinical Physicist's Guide to Statistical Analysis in Medical Physics Articles

Steven Sutlief, PhD DABR FAAPM

Banner MD Anderson Cancer Center

I prepared a detailed set of worked examples to show how these tests can be implemented in Excel. Unfortunately, non-PDF files may not be sharable as handouts. If that is the case, look me up in the AAPM directory and send me a request.

1

Conflicts of Interest

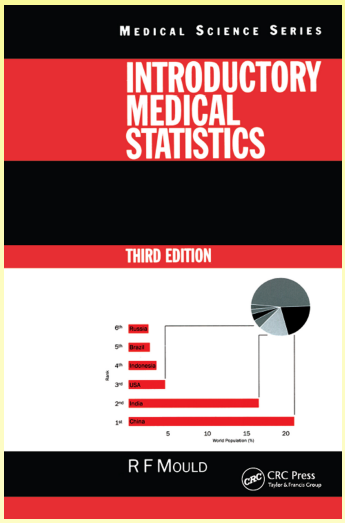
- I have no conflicts of interest regarding this presentation.

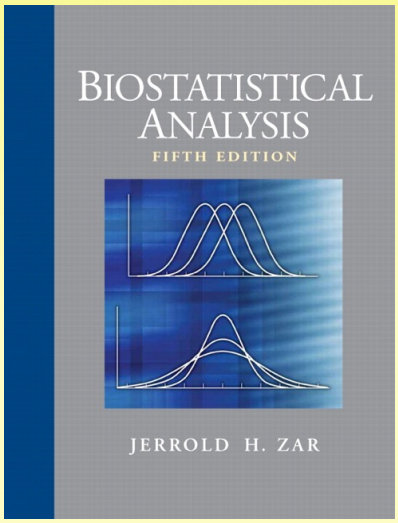
2

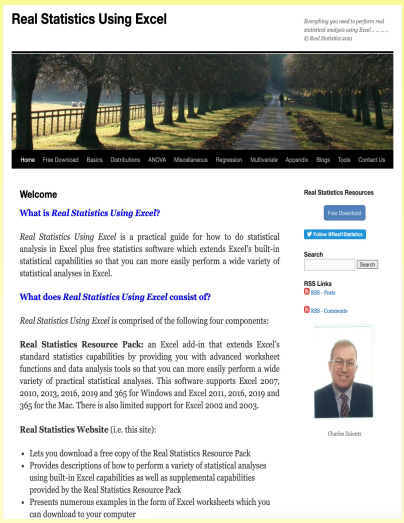
Educational Objectives

- Assess whether a research manuscript's use of a particular test is meaningful.
- Interpret the significance of the reported statistical findings.
- Select the appropriate tool for one's own research projects and estimate the sample size required to achieve the desired statistical power.

Resources Used for this Talk







Survey of Use of Statistical Tests

5

Survey of General US Journals

- Medical Physics (MP, free to AAPM members)
  - "publishes research concerned with the application of physics and mathematics to the solution of problems in medicine and human biology."
- Journal of the American College of Medical Physics (JACMP, Open Access)
  - "publishes papers that will help clinical medical physicists and other health professionals perform their responsibilities more effectively and efficiently for the increased benefit of the patient."

6

Statistics in Journals, Part 1

Search term	JACMP	MP
Total therapy articles	220	292
Total with terms found	59	40
Wilcoxon signed rank test	16	5
Pearson correlation	9	3
Student's t Test	7	1
Mann-Whitney	6	2
ANOVA	6	1

Statistics in Journals, Part 2

Search term	JACMP	MP
χ2 test	Unknown*	Unknown*
Shapiro Wilk	4	1
Bonferroni correction	3	0
Fisher's exact test	2	0
Kruskal–Wallis test	2	0
f-Test	2	0
Friedman test	2	0

*Hard to search for since usually represented by a Greek character.

Statistics in Journals, Packages

- R statistical package
- SPSS
- OriginPro
- JMP Software
- MATLAB
- Excel
- socscistatistics website

Review of Statistical Tests appearing in Articles about Medical Physics

Tests

- Tests for significance
  - Parametric tests
  - Non-parametric tests
- Tests for similarity
  - Dice, Cohen, ROC/AUC, Hausdorff Distance
- Tests for correlation
  - Linear regression, Pearson's coefficient

Can I assert that there is less than a 5% likelihood my data merely supports my claim by chance?

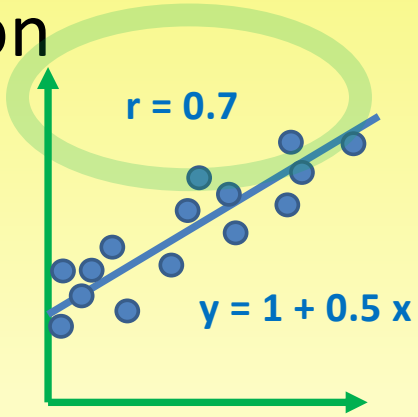
Can my AI software segment the liver as well as my expert users can?

Can my paired numerical data be modeled by a line? How well?

Tests for Sensitivity

- Dice similarity coefficient = 2 * |X∩Y| / (|X|+|Y|)
  - 0 = no agreement, 1 = complete agreement
- Receiver operator characteristics (ROC, AUC)
  - Plot of Sensitivity versus (1 - Specificity)
  - Sensitivity = true positive rate, Specificity = true neg. rate
- Also: Mean distance to agreement, Hausdorff distance
- Cohen's kappa: tests interobserver agreement (0 – 1).

Linear Regression



- Linear model for paired data
- Regression line = best fit line
- Regression coefficient = slope b
- Pearson's coefficient r = correlation (-1 to +1)
- Significance testing: t-Test, Spearman's test

Hypothesis Testing for Significance

- In general, we seek to rule out that the condition we are testing had no affect on the data (called the null hypothesis, H_0). **Examples:**
- H_0: Our data mean equals the population mean.
- H_0: Two data sets share the same mean.
- H_0: Three data sets share the same distribution.
- Generally desired significance level: $\mathbf{P \leq 0.05}$.

Significance Test Properties

- Normally distributed vs unknown distribution
- Single set, two set, three or more sets
- Paired data versus unpaired, vs frequency only
- Small data set, versus medium, versus large
- One tail versus two tail

15

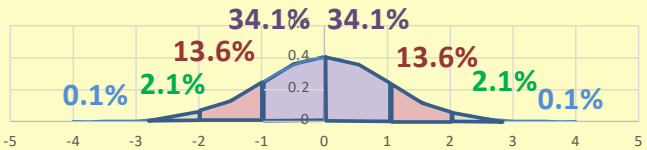
Taxonomy of Significance Tests

	One parameter, one set		Two parameters or sets		Three or more parameter or sets	
		Paired	Unpaired		Paired	Unpaired
Shaped like a normal distribution (Parametric, Numeral)	n ≤ 30: One Sample t Test n > 30: z Test	Paired t Test n any size	n any size: Unpaired t Test		Two-way ANOVA n any size	One-way ANOVA n any size
No assumed distribution shape (Non-parametric, Ordinal)	χ^2 test	Ranks: Wilcoxon signed rank test, Spearman, Sign Test, n ≤ 25	n ={1,…,5}: Wilcoxon rank sum test, Mann-Whitney U test		Friedman n any size	Kruskal–Wallis tests n > 20
		Finite Bins	N > 100 & n_i > 10: 2x2 contingency table N < 100 or n_i < 10 but N > 20: 2x2 w/ Yate's correction Else use this: Fisher's Exact Test			Generalized Contingency Table <20% of values < 5, All value ≥ 1

16

Normal Distribution

- Set of samples $\{x_1, \ldots, x_n\}$
- Mean $\bar{x} = (x_1 + \cdots + x_n)/n$
- Standard deviation $\sigma = \sqrt{s}$
- Variance $s = [(x_1 + \bar{x})^2 + (x_n + \bar{x})^2]/(n-1)$



A Baker's Dose of Significance Test Examples

Situation 1

- Case 1: Goodness of fit: We have a histogram of time-to-failure for MLC motors and we want to know whether it conforms to a log-normal shape.
- Case 2: 2x2 Contingency table: Is there an association between extended hours work and plan check errors?
- Null hypothesis: There is no difference between the observed and expected number in each data bin.
- Statistic: χ^2 Test

	All issues were caught	Chart issues Were missed
Check during hours	23	2
Check after hours	15	5

19

χ^2 Test

- The χ^2 test is non-parametric.
- For each data bin *i* we need the observed number O_i and the Expected number E_i.
- $\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$
- The χ^2 table look up depends also on degrees of freedom DF = n – 1 and desired sensitivity, $P \leq 0.05$.

20

JACPM Example of χ^2 Test

- Turchan *et al* investigated whether image registration and OAR delineation are often performed by medical physicists or dosimetrists without routine review by treating physicians.
- They used a Likert-type survey and assessed the binned data with a χ^2 test to obtain confidence limits on the frequency of the responses.

Turchan *et a*l. J. Appl. Clin. Med. Phys. 2020; 21:11:80–87.

Situation 2

- My data should have the same shape as the normal distribution and I want to know: does it's mean $\bar{x}$ differ from the hypothetical mean μ?
- Null hypothesis H_0: $\bar{x} = \mu$
- Statistic: Two-tailed one-sample *t*-Test

t-Test

- One-sample *t*-Test: inference based on a single sample mean x_m when the reference population from which the sample is drawn is known to have a mean of μ.
- $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$, s = st. dev. of the the sample

23

	A	B	C	D	E	F	G	H	I	J	K
1	Situation: Vendor claims the widgets will last 180 days on average.										
2	Is our lifetime data on ten widgets consistent with the claim?										
3								**Two-Tail t-Test**			
4	Null Hypothesis Ho:					Separate columns for the labled P-value					
5	The sample mean equals the claimed mean.				DF	0.2	0.1	0.05	0.02	0.01	0.001
6	180	x_cm (days) = Claimed mean			1	3.078	6.314	12.706	31.821	63.657	318.313
7					2	1.886	2.92	4.303	6.965	9.925	22.327
8	Widget #	Life (days)	(x_i - x_m)^2		3	1.638	2.353	3.182	4.541	5.841	10.215
9	1	197	136.89		4	1.533	2.132	2.776	3.747	4.604	7.173
10	2	183	5.29		5	1.476	2.015	2.571	3.365	4.032	5.893
11	3	184	1.69		6	1.44	1.943	2.447	3.143	3.707	5.208
12	4	187	2.89		7	1.415	1.895	2.365	2.998	3.499	4.782
13	5	191	32.49		8	1.397	1.86	2.306	2.896	3.355	4.499
14	6	192	44.89		9	1.383	1.833	2.262	2.821	3.25	4.296
15	7	179	39.69		10	1.372	1.812	2.228	2.764	3.169	4.143
16	8	184	1.69		11	1.363	1.796	2.201	2.718	3.106	4.024
17	9	174	127.69								
18	10	182	10.89								
19	Mean x_m:	185.3	6.70074623	<-- St.Dev. s_m = Sqrt(Sum((x_i - x_m)^2) / (N - 1))							
20	10	N = number of measurements									
21	9	Degrees of Freedom = N - 1									
22											
23	2.50122464	t-Test = (x_m - x_cm)/(s_m/N)									
24	2.262	t_critical(df=N-1,P=0.05)	(From the lookup table)								
25	Since t-Test > t_critical, we reject the null hypothesis that our sample shares the claimed mean lifetime.										

24

Situation 3

- My data should have the same shape as the normal distribution and I want to know: is it's mean $\bar{x}$ less than the hypothetical mean μ?
- Null hypothesis H_0: $\bar{x} \geq \mu$
- Statistic: One-tail one-sample *t*-Test

One-tail *t*-Test

- One-tail one-sample *t*-Test: inference based on a single sample mean x_m when the reference population from which the sample is drawn is known to have a mean of μ.
- Use significance $P \leq 0.025$ instead of $P \leq 0.05$
- $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$, s = st. dev. of the the sample

Situation 4

- My data has two sets representing different test conditions for the same participants, normally distributed, and I want to know: does the mean of the first set differ from the mean of the second set?
- Null hypothesis H_0: $\bar{x}_1 = \bar{x}_2$
- Statistic: Paired two-tail *t*-Test

Paired Two-tail *t*-Test

- Paired two-tail *t*-Test: inference based on a set of participants each receiving two different measurements, from which a mean value can be obtained from each measurement set: $\bar{x}_1$ and $\bar{x}_2$.
- The st.dev. of the differences must be computed.
- $t = \frac{\bar{x}_1 - \bar{x}_2}{s_{s1-s2}}$, $s_{s1-s2} = \frac{\sqrt{\sum(x_{1i} - x_{2i} - \overline{x_1 - x_2})}}{\sqrt{n-1}}$

JACMP Example of a *t*-Test

- Arbab *et al* developed a new margin formula that would cover the space occupied by an oropharyngeal clinical target volume (CTV).
- They used a *t*-Test to assign a P-value to the reduction in the superior PCM overlap in the base of tongue (BOT) lesions compared to tonsillar lesion due to a new rotational PTV.

Arbab *et al*. J Appl Clin Med Phys 2020; 21:11:172–178.

Situation 5

- My two data has two sets representing different test conditions for different participants, normally distributed, and I want to know: does the mean of the first set differ from the mean of the second set?
- Null hypothesis H_0: $\bar{x}_1 = \bar{x}_2$
- Statistic: Un-paired two-tail *t*-Test

Un-paired Two-tail *t* Test

- Paired two-tail *t*-Test: inference based different sets of participants each receiving a measurement. A mean value is obtained from each measurement set: $\bar{x}_1$ and $\bar{x}_2$. The pooled standard error of the differences must be computed.
- $t = \frac{\bar{x}_1 - \bar{x}_2}{SE_p}$, SE_p = pooled standard error

(Too complicated to fit on a slide)

Situation 6

- My data has two sets representing different test conditions for the same participants, no assumed distribution, and we want to know: is the second distribution different from the first distribution?
- Null hypothesis H_0: There is no difference between set 1 and set 2.
- Statistic: paired Wilcoxon rank sum test

Paired Wilcoxon Signed Ranks Test

- Compute the test differences between test 1 and test 2 for each participant.
- Rank the differences without regard to sign.
- Sum all the negative-sign ranks and separately sum all the positive-sign ranks.
- Wilcoxon test statistic is the larger of these two sums.

JACMP example:
Wilcoxon Signed Rank Test, Matched Pairs

- Akino *et al* looked at the differences between detector output factors and the corrected field output factor for different energies.
- For each beam energy, differences between the corrected and uncorrected output factors were compared using Wilcoxon signed rank test and statistical significance used a P value of <0.05.

Akino *et al.* J Appl Clin Med Phys 2020; 21:1: 78–87.

Situation 7

- My data has two sets representing different test conditions for different participants, no assumed distribution, and I want to know: is the second distribution different from the first distribution?
- Null hypothesis H_0: There is no difference between set 1 and set 2 distributions at the 0.05 level of significance.
- Statistic: Unpaired Wilcoxon rank sum test
- Statistic: Mann-Whitney U test

Unpaired Wilcoxon Rank Sum Test

- Rank all members from group 1 and group 2 together.
- Add the ranks of group 1 together and separately add the ranks of group 2 together.
- Wilcoxon test statistic is the greater of these two sums.

Mann-Whitney U Test

- This test statistic takes the result of the Wilcoxon rankings T_1 & T_2, but adjusts them to account for differences in samples sizes N_1 and N_2.

$$U = \{N_1 \cdot N_2\} + \{N_1(N_1 + 1)/2\} - T_1$$
$$U' = \{N_1 \cdot N_2\} + \{N_2(N_2 + 1)/2\} - T_2$$

- Take the smallest of these two values.

JACMP Example of Paired Wilcoxon

- Anusionwu *et al* used repeated measurements for several detectors to investigate the effect of experimental uncertainties on the small shifts observed in the PDD curves.
- They used a non-parametric (Mann–Whitney Wilcoxon) test to check whether the observed differences in the detector readings were statistically significant.

Anusionwu *et al.* J Appl Clin Med Phys 2020; 21:1: 117–126.

MP Example: Mann-Whitney U Test

- Babier *et al* developed a knowledge-based planning pipeline that does not use feature extraction from overlap-volume histograms.
- In their testing of the workflow, they evaluated how similar the KBP predictions were to their corresponding clinical plans.
- They used a one-sided Mann–Whitney U test to determine whether the 3D-dose predictions had the same or greater absolute error than the predictions from the other KBP models.

Babier et al, Med. Phys. 47 (2):297-306, February 2020.

39

Situation 8

- My data has more that two sets representing different test conditions, assumed normal distribution, and I want to know: are the means of the distributions the same?
- Null hypothesis H_0: $\mu_0 = \mu_1 = \mu_2 = \ldots$ at the 0.05 level of significance.
- Statistic: One-way ANOVA

40

One Way ANOVA

- This test statistic takes multiple data sets and reports the likelihood that they share the same mean μ.
- It does this by combining the means and variations within and between datasets to obtain an f-Statistic (F = mean square between groups/mean square within groups).

41

Situation 9

- My data has more that two sets of data and within each I am performing a paired set of measurements (factor-A and factor-B). I want to know: are the distributions the same?
- Null hypothesis 1: H_0: All of the factor-A means are equal.
- Null hypothesis 2: H_0: All of the factor-B means are equal.
- Null hypothesis 3: H_0: There is no interaction between factor-A and factor-B.
- Statistic: Two-way ANOVA

42

Two Way ANOVA

- This test statistic takes multiple data sets and reports the likelihood that they share the same mean m_A for factor-A, the same mean m_B for factor-B, and that there is no interaction between the factors.
- Like the one-way ANOVA, it does this by combining the means and variations within and between datasets to obtain an f-Statistic. It is more complicated though, because the effect of variation in the factor-A data and the factor-B data must be handled separately.

Situation 10

- Like the one-way ANOVA situation, my data has more that two sets representing different test conditions, but now there is no assumed distribution. I now want to know: are the medians of the distributions the same?
- Null hypothesis: $median_0 = median_1 = median_2 = \ldots$
- Statistic: Kruskall-Wallis Test

Kruskal–Wallis Test

- Combine all data from each data set together in one rank order. Compute the rank sums R_i for each data set *i*.
- Apply the Kruskal-Wallis statistic H:

$$H = \frac{12}{N(N+1)}\left(\sum \frac{R_i^2}{n_i}\right) - 3(N-1)$$

- If the null hypothesis were true, *H* would follow a χ^2 distribution where the degrees of freedom is one less than the number of samples. Refer to a χ^2 table for significance.

JACMP Example of the Kruskal–Wallis Test

- Chen *et al* looked at the influence of MLC leaf speed on the quality of VMAT plans.
- They used the nonparametric Kruskal–Wallis test for comparing plan quality metrics in data with seven different leaf speeds.

Chen *et al*. J Appl Clin Med Phys 2020; 21:11:37–47.

Situation 11

- This situation is like that for the Wilcoxon signed rank test, my data may not be normally distributed, but this time it has three or more measures: I have a set of samples, each with data for each of the measures. I want to test for whether there is a preferred order among the measures.
- Null hypothesis: There is no difference between the measures.
- Statistic: Friedman Test

Friedman Test

- This test is very similar to the Kruskal-Wallis test, although now we are ranking a set of alternatives *k* in *K* for each subject *j* in *J*.
- Combine all data from each data set together in one rank order. Compute the rank sums R_k for each data set *k*.
- Apply the Friedman's statistic T:

$$T = \frac{12}{JK(K+1)}\left(\sum R_k^2\right) - 3J(N-1)$$

- If the null hypothesis were true, *T* would flow a χ^2 distribution where the degrees of freedom is one less than the number of samples. Refer to a χ^2 table for significance.

JACMP Example of Friedman Test

- Casati *et al* created an auto-contouring atlas methodology for MiM.
- They used the Friedman test to assess differences between three degrees of smoothing factor using the Dice similarity coefficient and mean distance to agreement.

Casati *et al.* J Appl Clin Med Phys 2020; 21:12:219–230.

49

Situation 12

	# linacs $\leq$ 4	# linacs > 4	Sums
Onsite engineer	a = 3	b = 6	a + b = 9
Vendor support	c = 12	d = 4	c + d = 16
Sums	a + c = 15	b + d = 10	N = 25

- We have a 2x2 contingency table with small values and want to test for a correlation. E.g., number of department linacs and staff engineers.
- Null hypothesis: No association between the two parameters.
- Statistical tool: Fisher Exact Probability Test

50

Fisher Exact Probability Test

- Probability of the observed distribution of frequencies under the null hypothesis H_0:

$$P = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{N!\,a!\,b!\,c!\,d!}$$

- By working out the probabilities for our table and those more extreme, we can determine that overall likelihood our table arose by chance, when H_0 holds.

JACMP Example of Fisher's Exact Test

- Berry *et al* developed and implemented an automated plan check tool.
- They used a two-tailed 2-by-2 contingency table using Fisher's exact test to determine whether the change in the number of un-related and related errors before and after introduction of the tool was statistically significant.

Berry *et al*. J Appl Clin Med Phys 2020; 21:4:51–58

Situation 13

- We have a set of tests, each of which yields its own P-value for significance. We want to know what P-value these tests must remain below to ensure the overall P-value threshold is not exceeded.
- Statistical Tool: Bonferroni Correction
- *When possible, ANOVA or Friedman are preferable.*

Bonferroni Correction

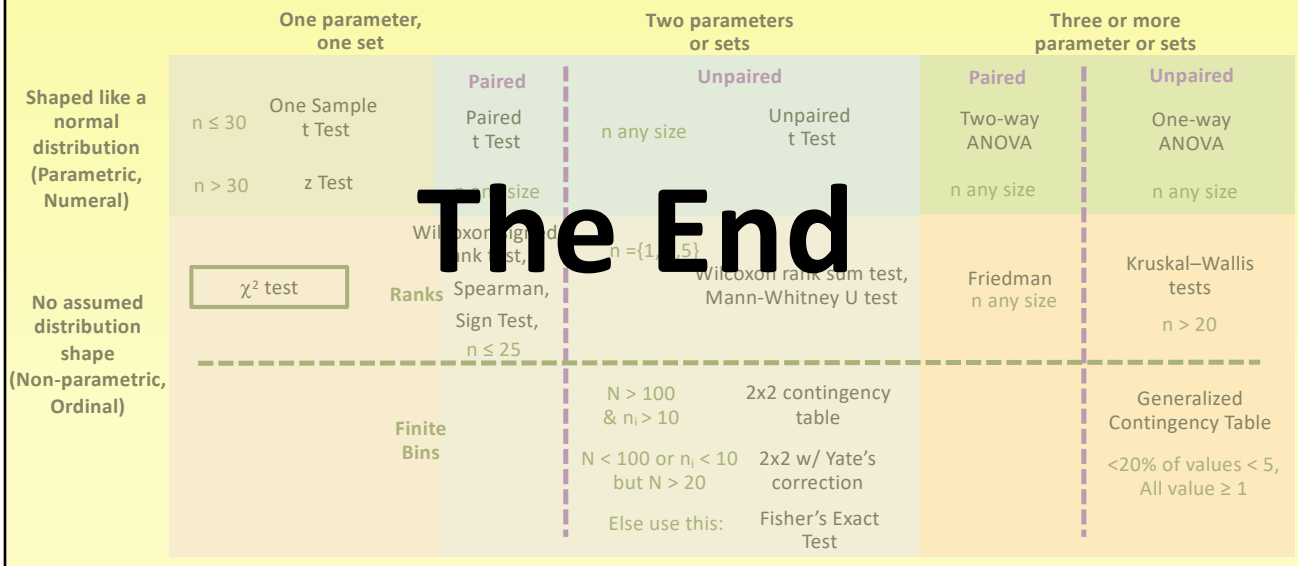
- The Bonferroni method depends on the overall false positive rate we are trying to achieve across a collection of *k* independent subgroups.
- We usually take this to be $\alpha = 0.05$, corresponding to our $P \leq 0.05$ significance criteria.
- Single-test cutoff: $\alpha' = 1 - \sqrt[k]{1-\alpha}$ ($\alpha' \approx 0.01$ for k=5)

JACMP Example of Bonferroni Correction

- Xiao *et al* studied the dosimetric impact of daily positioning variations measured with CBCT on whole-breast radiotherapy patients treated prone.
- They used paired Student's t tests at $\alpha = 0.01$ to compare structure dose metrics across plans after Bonferroni correction for multiple testing resulting in a $P < \alpha' = 0.002$ (i.e., $P < 0.01/5$).

Xiao *et al*. J Appl Clin Med Phys 2020; 21:12:146–154.

55



The End

56