Use and Abuse of Common Statistics in Radiological Physics

Zacariah Labby, Ph.D., DABR
Department of Human Oncology
University of Wisconsin - Madison

Conflicts of Interest

None to disclose

Outline

Hypotheses

One-Sample Statistics

Two-Sample Statistics

Statistics of Agreement

Statistics of Time Data

A plug for "R"…

R is a free software package for data analysis and is very common in the statistics community.

Good text for learning R and basic stats:

Statistics: An Introduction using R by Michael J. Crawley, published 2005 by John Wiley & Sons, Ltd

Hypotheses

- **A good hypothesis is a falsifiable hypothesis**
- Hypothesis 1: There are cancer cells in my body.
- Hypothesis 2: There are no cancer cells in my body.
- How can I reject these hypotheses?
- How will this apply to Null Hypotheses?

One-Sample Inference

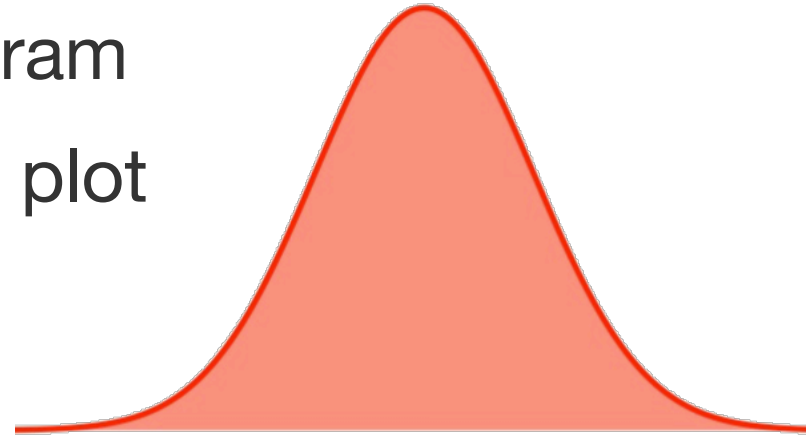
Test: Is the "middle" of our sample consistent with an assumed value?

- **Parametric**: Student's *t*-test
- **Non-Parametric**: Wilcoxon signed-rank test
- **"Parametric"** refers to appropriateness of assumed parameterization (e.g., Normal)
- Student's *t* distribution is appropriate for (approximately) Normal sampled data

Normality

Simple methods to analyze normality

- Look for bell-curve histogram
- Look at Quantile-Quantile plot



Student's *t* distribution is appropriate for a sampled Normal distribution (N < 30)

- More data? Use either *t*- or Z-tests

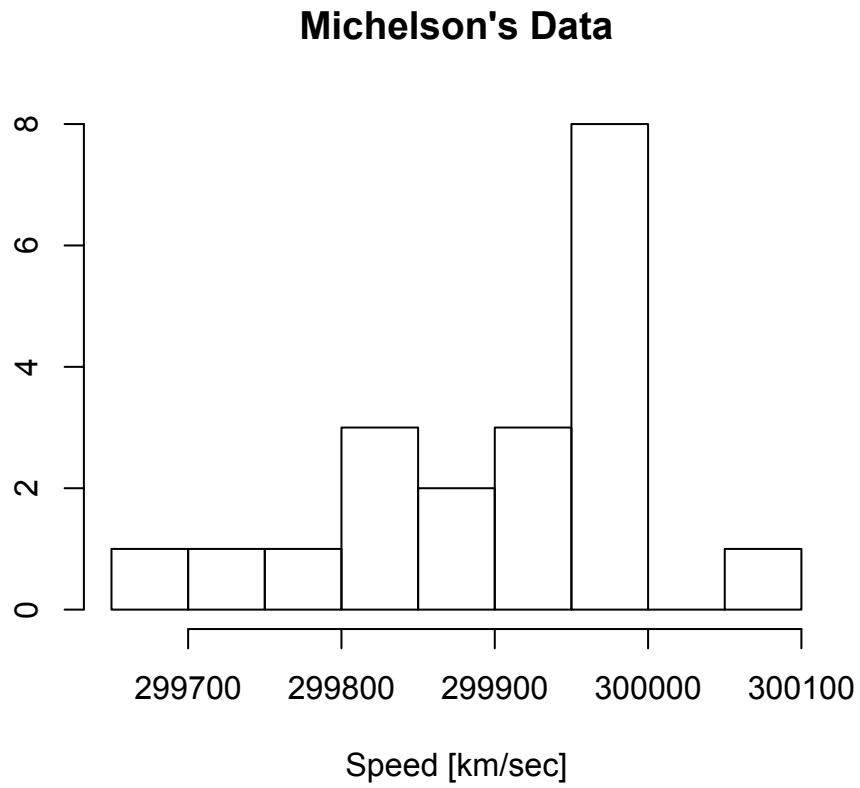
One-Sample Inference

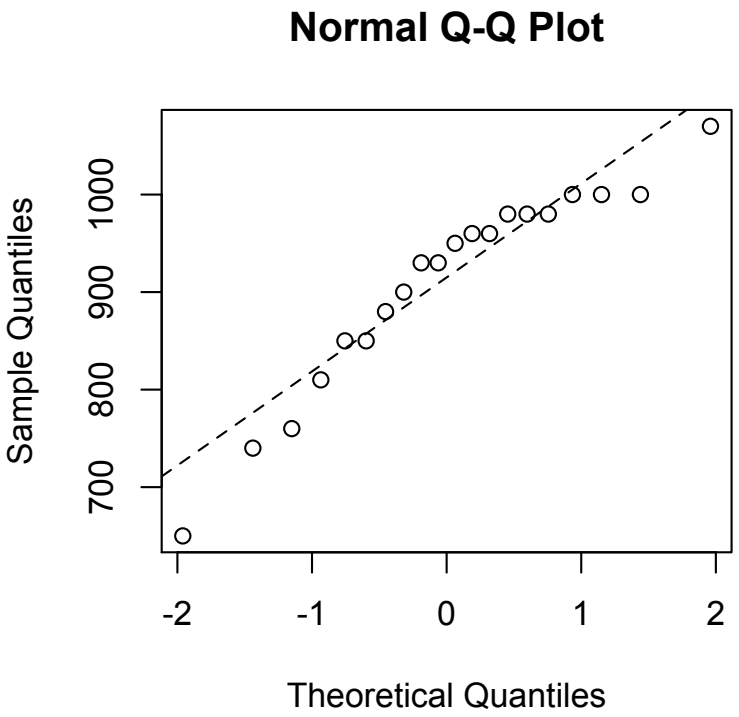
Example: Albert Michelson's data on the speed of light (late 1870's)

- Collection of measured speeds using rotating and fixed mirrors
- Is data consistent with prior knowledge at the time? (299,990 km/sec at the time of measurements)
- Null Hypothesis: Mean of data (speed of light) is equal to 299 990 km/sec

One-Sample Inference

Which test is appropriate?





One-Sample Inference

Which test is appropriate?

Not Normally distributed, so use Wilcoxon signed-rank test against a value of 299990

```
wilcox.test(LightSpeedData,mu=299990)
```

p-value: probability of finding this particular data if the Null Hypothesis were true

$p = 0.00213$, so we'll probably reject. Speed is "significantly" different from prior value.

Two-Sample Tests

Comparing two

- Means
- Proportions
- Distributions

Two-Sample Tests – Comparing Means

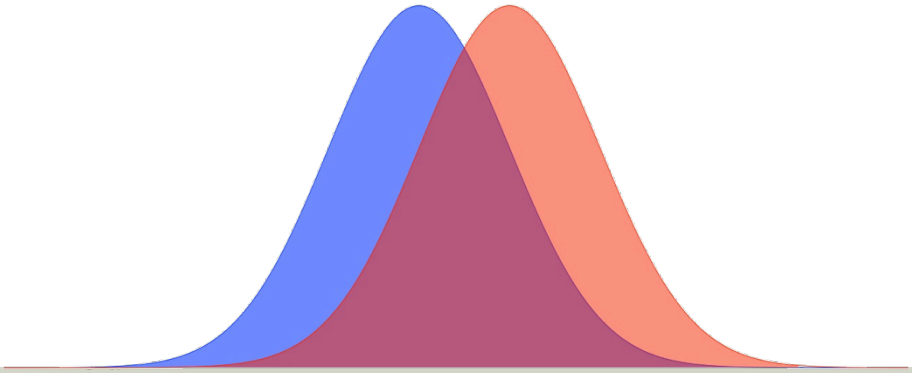
Null Hypothesis: means are equal

- Alternatives: Not Equal, Greater, or Less
- Choose “two-sided” without *a priori* reason
- Choose “one-sided” if that’s all you care about

Two-Sample Tests – Comparing Means

Independence of samples ("Unpaired")

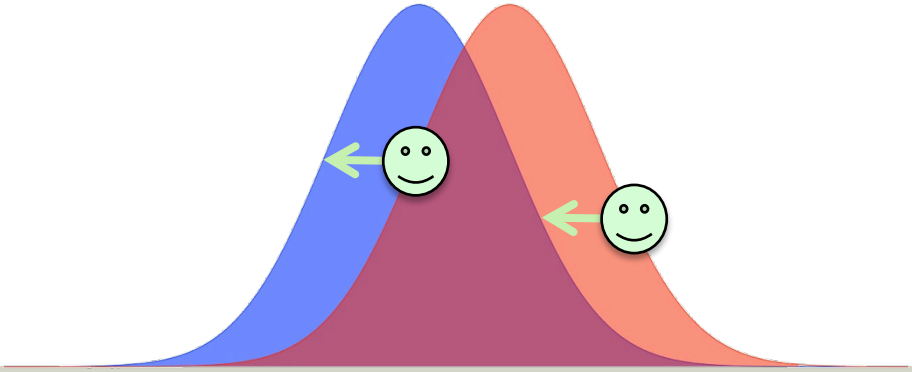
- Are the two samples linked?
- Before and After intervention?
  - Example: Lung perfusion before and after RT
- Pairing adds substantial power
- Paired test equivalent to one-sample (x_1-x_2) test



Two-Sample Tests – Comparing Means

Independence of samples ("Unpaired")

- Are the two samples linked?
- Before and After intervention?
  - Example: Lung perfusion before and after RT
- Pairing adds substantial power
- Paired test equivalent to one-sample (x_1-x_2) test



Two-Sample Tests – Comparing Means

Parametric test: Student's *t*-test

- Both samples are approximately Normal
- Paired or Independent tests

Non-parametric test: Wilcoxon tests

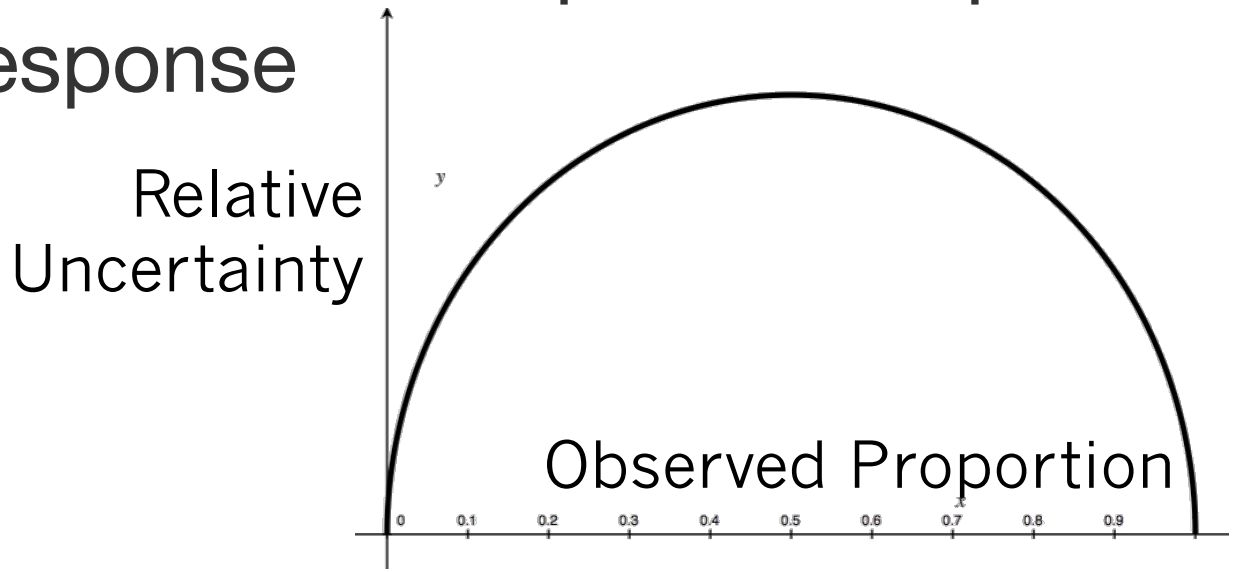
- No assumed distribution; more power when not Normal
- Signed Rank test for paired data
- Rank-Sum test for independent samples
- Data needs to be at least "ordinal"

Two-Sample Tests - Proportions

Quantifying Probability of Events

- Information in both Total Count and Responses
- Variability decreases with increasing counts

Standard Deviation has specific shape for binomial response



Two-Sample Tests - Proportions

Specific Proportion Tests exist

- Quantify the observed proportion
- Report the confidence interval on the proportion
  - Model-based or Fischer's Exact interval (better for small *N* or "extreme" proportions)
- Test the observed proportion against a null hypothesis
- Possible example: proportion of Radiation Workers exceeding Occupational Exposure ALARA levels

Two-Sample Tests – Distributions

Tests of means are most common, but…

Can test for equal variances or "scales"

- *F* test for Normal distributions
- Ansari-Bradley test for non-Normal data

Could be used on its own (precision of data)

- Example: new daily QA device, precision vs. accuracy

Could be used to give Power to *t*-tests

- With equal variance, the test is more efficient

Two-Sample Tests – Distributions

Broad Question: Are two distributions the same?

Not just mean, not just variance…everything.

Kolmogorov-Smirnov test

- Data needs to be continuous

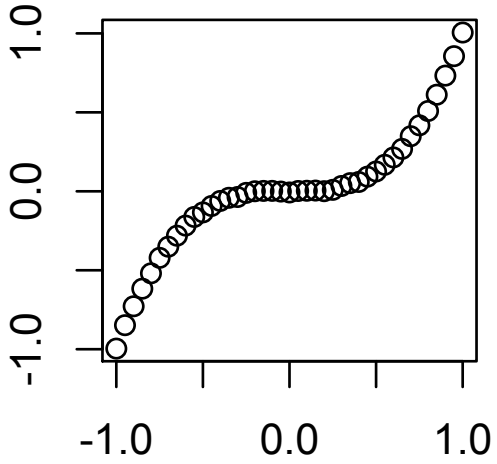
Null Hypothesis: Distributions are identical

- Alternative: Not identical

Statistics of Agreement

For Continuous data: Correlation

- Parametric: Pearson's *r*
  - Parameterization: Straight Line
- Non-Parametric: Spearman's ρ (rank correlation)
  - If one goes up, does the other go up?



r=0.92

ρ=0.99

Statistics of Agreement

For Categorical data, we can assess "reliability" between raters

- Example: 30 image sets, two observers, rating tumor visibility as "good," "moderate," or "poor"
- How well do the observers agree?

Use κ (kappa) statistics

kappa: Extent of agreement between observers beyond that expected by chance

$\kappa=1$, perfect agreement; $\kappa \leq 0$, no agreement

Statistics of Time Data

For time-to-event data, use special Survival Analysis statistics

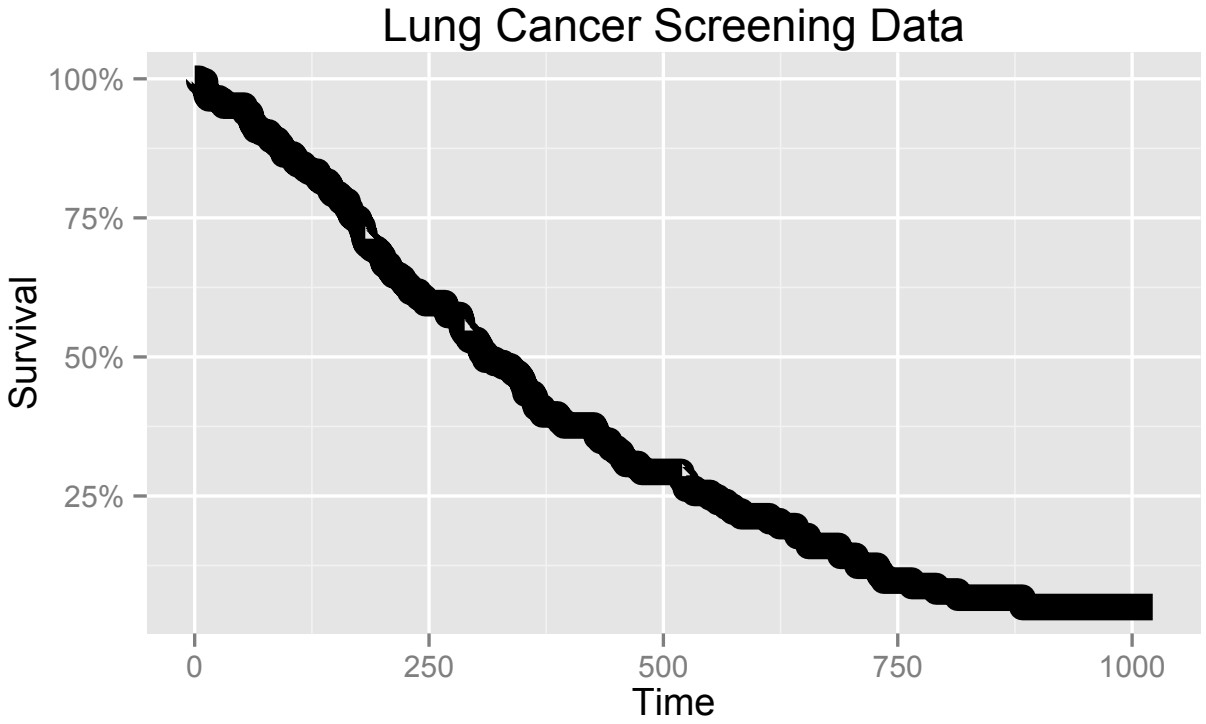
Many facets

- Partly proportional data
- Partly non-parametric data
- Data changes over time

Survival data can also have “censoring”

- Patients lost to follow-up

Statistics of Time Data



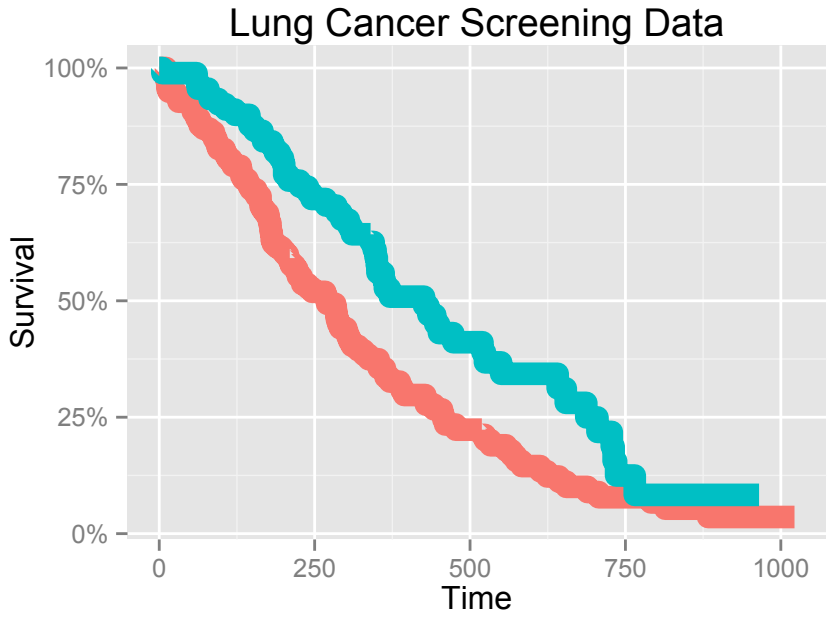
"Kaplan-Meier" curve is the estimate of survival

Can extract statistics for standard metrics, e.g., median survival

Statistics of Time Data

Test for differences between groups: Log-Rank test

- Null hypothesis: equal hazard rates (patients die at the same frequency between groups)



***p* = 0.0013**
Reject null hypothesis

Use your Biostatisticians

Many large centers have at least one biostatistician on staff

In many centers, free consultations for

- Experimental design
- Simple clinical trials
- Data analysis questions

Paid services will often prevent headaches and lost costs for rework and rejected papers

WISCONSIN

UNIVERSITY OF WISCONSIN–MADISON