Nonparametric Statistics

STA6507

Dr. A Cohen - acohen@uwf.edu

Mathematics and Statistics Department
University of West Florida

2021

This course covers nonparametric statistical methods.

- Class Meetings MW @4:00 PM CT via Zoom
- R/RStudio will be used
- Assignments (4-6)
- Take Home Exams (2)

If you catch any typos or inconsistencies in the slides please refer them to me at acohen@uwf.edu

Introduction to Statistics and Probability

What is Statistics?

The next time you read a newspaper, look for items such as the following:

- 53% of the people surveyed believe that the president is doing good job
- The average selling price of a new house is $74, 500
- The unemployment level is 4.9%

The numerical facts or data in the news items (53%, $74, 500, 4.9%) commonly are referred to as statistics.
In common, everyday usage, the term statistics refers to **numerical facts** or **data**.

What is Statistics?

The field of statistics involves much more than simply the computation and presentation of numerical data.
In a broad sense the subject of statistics involves the study of:

- how data are **collected**
- how they are **analyzed**
- how the results are **interpreted**
- Data **visualization**

A major reason for collecting data, analyzing, and interpreting data is to provide engineers, managers, public, other researchers, with the information needed to make **effective decisions**.

What is Statistics?

Descriptive In many statistics studies we are interested only in summarizing a set of data in order to present it in a more convenient or more easily interpreted form.

- **Central Tendency**: mean, median
- **Dispersion**: variance, standard deviation, IQR
- **Distribution**: histogram, Boxplot, scatter

Inferential Much of statistics is concerned with analyzing sample data in order to learn about characteristics of a population (parameters).

- **Probability distributions**
- **Estimation**
- **Hypothesis Testing**
- **Confidence Intervals**

What is Statistics?

Descriptive In many statistics studies we are interested only in summarizing a set of data in order to present it in a more convenient or more easily interpreted form.

- **Central Tendency**: mean, median
- **Dispersion**: variance, standard deviation, IQR
- **Distribution**: histogram, Boxplot, scatter

Inferential Much of statistics is concerned with analyzing sample data in order to learn about characteristics of a population (parameters).

- **Probability distributions**
- **Estimation**
- **Hypothesis Testing**
- **Confidence Intervals**

This course covers Nonparametric Statistics.
What is the nonparametric Statistics?

What is Statistics?

Nonparametric Statistics?

Nonparametric statistics or distribution-free methods make no assumptions about the underlying probability distributions. On the other side, **Parametric statistics** depends on knowing the population distribution (e.g. normal).

Examples

Most parametric procedures assume the normality or other parametric distribution of the response variable such as: ANOVA, t-tests, linear regression models.

Advantages Probability statements obtained from most nonparametric statistical tests are **exact probabilities**.

If a **sample size as small** as n=3 is used, there is no alternative to using a nonparametric statistical tests unless the population distribution is know exactly.

The nonparametric tests are only slightly **less efficient than their normal theory** competitors when the underlying populations are normal.

Nonparametric methods are relatively **insensitive to outliers.**

Nonparametric methods are available to treat data which are **ordinal and nominal** scales.

Disadvantages

Nonparametric tests can be **less powerful** than the parametric tests under the required assumptions (e.g. normal distribution).

Nonparametric procedures **lose information**, because they only use the orders, signs, and differences while eliminating the real values of the data.

Experiment

An experiment is the process of following a well defined set of rules, where the result of following those rules is not known prior to the experiment.

Examples

Tossing of a coin, rolling a die.

All possible outcomes of an experiment (**sample points**) constitute the **sample space**.

Rules that help us **count!** the sample space size and therefore find probabilities.

Rule 1

If an experiment consists of **n trials** where each trial may result in one **k possible outcomes**, there are $\mathbf{k^n}$ **possible outcomes** of the entire experiment.

Examples

- Tossing of a coin: 2^1
- Tossing of a coin 5 times: $2^5 = 32$
- Rolling a die: 6^1

Rule 2

There are **n**! ways of arranging n distinguishable objects.

Examples

- n plastic chips in a box.
- How many ways can we arrange the letters *A*, *B*, and *C*?

Rule 3

If a group of n objects is composed of n_1 identical objects of type 1, n_2 identical objects of type 2,..., n_r identical objects of type r, the number of distinguishable arrangements int a row, denoted by:

$$\frac{n!}{n_1!n_2!\ldots n_r!}$$

If a group of n objects is composed of k identical objects of one kind and the remaining (n-k) objects are identical of a second kind, the number of distinguishable arrangements of n objects into a row is given by:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Examples

- Find the probability of a football team to win at least 7 games in a 8-game season.

Sample space is the collection of all possible different outcomes of an experiment.

Sample point is a possible outcome of the experiment.

An event is any set of point in the sample space.

If we repeat the experiment under fairly uniform conditions, the relative frequency of the occurrence of the point represents an approximation to the probability of that point.
These probabilities may be any number between 0 and 1.

Probability

If A is an event associated with an experiment, and if n_A represents the number of times A occurs in n independent repetitions of the experiment, **the probability of the event** A, denoted by $P(A)$, is given by:

$$P(A) = \lim_{n\to\infty} \frac{n_A}{n} \tag{1}$$

Which is read *"the limit of the ratio of the number of times A occurs to the number of times the experiment is repeated, as the number of repetitions approaches infinity"*.

Probability Concepts: Properties

- If A is an event, then $0 \leq P(A) \leq 1$.
- Let the whole sample space be denoted S. Then $P(S) = 1$.
- $P(A) = 1 - P(\overline{A})$
- If A and B are events, then $P(A \cup B) = P(A) + P(B) - P(AB)$.
- If A and B are independent events, then $P(AB) = P(A)P(B)$.
- If events A and B are mutually exclusive (or disjoint), then $P(A \cap B) = 0$.
- The conditional probability of A given B is $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$

Random variables

Most of the statistical studies are interested in numerical values such as the number of days it rains; the number of patients suffering from a side effect and so forth. We can define **many random variables of interest** associated with a sample space.

Random variable

A random variable is a function that assigns real numbers to the points in a sample space.

Examples

Tossing a coin twice: HH, HT, TH and TT.
Define X the number of heads then X can take on $0, 1, 2$.

Random variables & probability function

Random variables can be either Continuous or Discrete.

Discrete can take on a countable number of possible outcomes.

Continuous can take on any value in an interval.

Probability function

A Probability function is a function that assigns probabilities to each sample point (each value of a random variable X).

Discrete Probability Mass Function - PMF

Continuous Probability Density Function - PDF

Another probability function is the **Cumulative Distribution Function** - CDF, denoted by $F(X \leq x)$.

- **Binomial distribution**: $X \sim Binom(n, p)$; where n is number of trials and p is probability of having a “success" outcome in each trial.
- **Normal distribution**: $X \sim N(\mu, \sigma^2)$; where μ is the population mean and σ^2 is the population variance.
- **Chi-squared distribution**: $X \sim \chi^2(K)$; where K is the degrees of freedom.
- **F-distribution**: $X \sim F(n, m)$; where n and m are the degrees of freedom.

Random variables & probability function

Sometimes we characterize a random variable (r.v.) X using:

Expected value $E(X) = \sum_{\forall x} xP(X = x)$

Variance $Var(X) = E(X^2) - [E(X)]^2$

Quantile The points dividing the distribution of X into equal intervals. The number x_p for a given value of p is called the p^{th} quantile of the r.v. X if

$$P(X < x_p) \leq p \qquad \textbf{and}$$
$$P(X \leq x_p) \geq p$$

Quantiles

2-quantile = Median; 4-quantile=quartile; 100-quantile = percentile.

Bootstrap for Confidence Intervals

- One of the principal goals of a statistic is to estimate an unknown population parameter. Commonly, estimation is the process of finding values/estimates/approximations of the population parameters based on the sample.
- An estimator is the method that provides a **point estimate** of the unknown population parameter.
- An interval estimate can be used to provide a range of values of the unknown parameter. A **confidence interval** consists of a range of values of the unknown population parameter with a certain confidence level.
- **The confidence interval is random.**

For example, if we would like to find the confidence interval for the population mean μ, we have various situations:

Normal data: $\bar{x} \pm z_{1-\alpha/2}(\frac{\sigma}{\sqrt{n}})$

Non-Normal data, large sample size $n \geq 30$: $\bar{x} \pm z_{1-\alpha/2}(\frac{\sigma}{\sqrt{n}})$

Normal data, σ is unknown: $\bar{x} \pm t_{1-\alpha/2}(\frac{s}{\sqrt{n}})$

Non-Normal data, large sample $n \geq 30$: $\bar{x} \pm t_{1-\alpha/2}(\frac{s}{\sqrt{n}})$ or $\bar{x} \pm z_{1-\alpha/2}(\frac{s}{\sqrt{n}})$, they will give similar results.

Non-Normal data, small sample $n < 30$: Nonparametric confidence interval (Bootstrap technique)

Bootstrap for Confidence Intervals

Why?

- When the **population distribution is unknown**.
- The **derivations are too difficult** to find the variance of a statistic (e.g. the sample variance)
- The **sample size is small**, say n=3

The bootstrap technique can be used to find confidence intervals for any population parameter.

How?

The bootstrap methods proceed with resampling (with replacement) from the data (original sample) in order to generate a large number of bootstrap samples.

Example

Suppose we have 10 data points: $1, 1, 2, 3, 3, 3, 5, 4, 7, 7$. We view this as a sample taken from an underlying distribution. A resampled data (with replacement) of 5 observations from these 10 data points would look like this $1, 1, 3, 7, 1$.

Principal

Suppose we have $X_1, X_2, \ldots, X_n$ drawn from a distribution F. A bootstrap sample $X_1^*, X_2^*, \ldots, X_n^*$ is a resampled data of the same size. You can think about the bootstrap sample as a sample drawn from the empirical distribution F^*. **The bootstrap setup will follow:**

- $X_1, X_2, \ldots, X_n$ is a data sample drawn from a distribution F

Principal

Suppose we have $X_1, X_2, \ldots, X_n$ drawn from a distribution F. A bootstrap sample $X_1^*, X_2^*, \ldots, X_n^*$ is a resampled data of the same size. You can think about the bootstrap sample as a sample drawn from the empirical distribution F^*. **The bootstrap setup will follow:**

- $X_1, X_2, \ldots, X_n$ is a data sample drawn from a distribution F
- θ is a statistic calculated from the sample

Principal

Suppose we have $X_1, X_2, \ldots, X_n$ drawn from a distribution F. A bootstrap sample $X_1^*, X_2^*, \ldots, X_n^*$ is a resampled data of the same size. You can think about the bootstrap sample as a sample drawn from the empirical distribution F^*. **The bootstrap setup will follow:**

- $X_1, X_2, \ldots, X_n$ is a data sample drawn from a distribution F
- θ is a statistic calculated from the sample
- F^* is the empirical distribution of the data

Principal

Suppose we have $X_1, X_2, \ldots, X_n$ drawn from a distribution F. A bootstrap sample $X_1^*, X_2^*, \ldots, X_n^*$ is a resampled data of the same size. You can think about the bootstrap sample as a sample drawn from the empirical distribution F^*. **The bootstrap setup will follow:**

- $X_1, X_2, \ldots, X_n$ is a data sample drawn from a distribution F
- θ is a statistic calculated from the sample
- F^* is the empirical distribution of the data (resampling)
- $X_1^*, X_2^*, \ldots, X_n^*$ is a resampled data of the same size

Principal

Suppose we have $X_1, X_2, \ldots, X_n$ drawn from a distribution F. A bootstrap sample $X_1^*, X_2^*, \ldots, X_n^*$ is a resampled data of the same size. You can think about the bootstrap sample as a sample drawn from the empirical distribution F^*. **The bootstrap setup will follow:**

- $X_1, X_2, \ldots, X_n$ is a data sample drawn from a distribution F
- θ is a statistic calculated from the sample
- F^* is the empirical distribution of the data (resampling)
- $X_1^*, X_2^*, \ldots, X_n^*$ is a resampled data of the same size
- θ^* is a statistic calculated from the bootstrap sample

The core concepts of the bootstrap technique

1. $F^* \Leftrightarrow F$
2. The variation of θ is well approximated by the variation of θ^*
3. The bootstrap is based on the **law of large numbers** (with enough data the empirical distribution will be a good approximation of the true distribution)
4. It is worth to say that resampling will not improve our point estimate.

They are different ways to construct the bootstrap confidence intervals.

Percentile method

It involves creating many bootstrap samples (say, 5000), and calculating the observed statistic θ^* for each bootstrap sample. Then a 95% confidence interval for θ would be:

$$\theta^*_{0.025} \leq \theta \leq \theta^*_{0.975} \tag{2}$$

The Residual method

The residual method is based on the distribution of the residuals from the original estimate $\hat{\theta}$. We create large number of bootstrap samples and calculate the observed statistic θ^* for each bootstrap sample. We then compute $e^* = \theta^* - \hat{\theta}$. Then a 95% confidence interval for θ would be:

$$\hat{\theta} - e^*_{0.975} \leq \theta \leq \hat{\theta} - e^*_{0.025} \quad (3)$$

There is another method called **BCa method (Bias-Corrected and accelerated)**, this method chooses the lower and the upper bounds to make the interval median unbiased and adjust for skewness.

R: The package **boot** in R has several functions to perform the bootstrap technique.

Definition

Hypothesis Testing is the process of inferring from a sample whether or not a given statement about the population appears to be true.

Examples

- Whether or not the treatment is effective?
- Method 1 is better than method 2?
- Process is statistically in-control?

Hypothesis Testing

- **Null hypothesis H_0:** it is usually formulated for the express purpose of being rejected.
    - No differences $(=; \leq; \geq)$
    - The process is in-control
- **Alternative Hypothesis** H_1: If H_0 is rejected, the H_1 may be accepted. It's the statement that the experimenter would like to prove.
    - There are differences $(\neq; <; >)$
    - The process is out-of-control
    - The quality of the product or service is unsatisfactory

- **The test statistic** is chosen to be sensitive to the difference between the null hypothesis and the alternative hypothesis.
  - A confidence interval is the inversion of a hypothesis test in that the confidence interval is the collection of the null hypotheses that are not rejected by the data
  - A powerful hypothesis test relates to a short confidence interval.
- **Level of significance** (α): Since the level of significance goes into the determination of whether H_0 is or is not rejected, the requirement of objectivity demands that α **be set in advance**.
  - The level should be determined by our estimate regarding the importance of our findings
  - The type II error β which is inversely proportional to α (type 1 error). If we would like to decrease α mechanically we will increase the type II error, β.

Hypothesis Testing

- **The Null distribution** is the distribution of the test statistic when H_0 is true. This defines the rejection region along with the level of significance.
- **The Power**: The probability of rejecting H_0 when it is in fact false. $Power = 1 - \beta = P(Rejecting H_0 | H_0 is FALSE)$.
- **P-value** is the probability, computed assuming that H_0 is true, that the test statistic would take a value as extreme or more extreme than that actually observed.
- **P-value** is a random variable.
- Accepting or Failing to Reject H_0 does not mean that the data prove the null hypothesis to be true.

Hypothesis Testing

	Decision	
	Fail to Reject H_0	Reject H_0
H_0 is TRUE	Correct decision (1-α)	Type I error - α
H_0 is FALSE	Type II error - β	Correct decision (1-β)

Hypothesis Testing: Properties

Once the hypotheses are formulated, there are generally several hypothesis tests available for testing the null hypothesis. We consider:

- Are **the assumptions** of the test met?
- **Unbiased** test: test should be more likely to reject H_0 when H_0 is false than when H_0 is true. $\boldsymbol{power = 1 - \beta > \alpha}$
- **Consistent** test: $\lim_{n \to \infty} Power \to 1$
- **Conservative** test: the actual level of significance is smaller than the stated level of significance.

Relative Efficiency

- The concept is concerned with the amount of **increase in sample size** which is necessary to make a test, say B, as powerful as a test A.
- The Relative Efficiency is the ratio

$$n_2/n_1$$

where n_1 and n_2 are the sample sizes of two tests T_1 and T_2, respectively, in order to have the same power under the same level of significance. **We read *the relative efficiency of T_1 to T_2* or the efficiency of T1 relative to T2.**
- The Asymptotic Relative Efficiency (A.R.E) is the limit of n_2/n_1 when n_1 approaches infinity.

Hypothesis Testing: Properties

Test 1	Test 2	A.R.E.	Assumptions
Sign Test	Paired t test	0.637	diff. are i.i.d normal
Sign Test	Wilcoxon signed test	0.667	diff. are i.i.d normal
Sign Test	Paired t test or Wilcoxon signed	0.333	diff. are i.i.d uniform (light-t
Sign Test	Paired t test	2	diff. are i.i.d Laplace (heavy-
Sign Test	Wilcoxon signed	1.333	diff. are i.i.d Laplace (heavy-
Cox-Stuart Trend	t test on regression coefficient	0.78	normal
Cox-Stuart Trend	Spearman's or Kendall's rank	0.79	normal
Cox-Stuart Trend	Spearman's or Kendall's rank	0.79	normal
The Median test	one-way ANOVA	0.64	Normal
The Median test	one-way ANOVA	2	Laplace (heavy-tailed)
Mann-Whitney Test	Two-Sample t test	0.955	Normal
Mann-Whitney Test	Two-Sample t test	1	Uniform
Mann-Whitney Test	Two-Sample t test	1.5	Laplace
Mann-Whitney Test	The Median test	1.5	Normal
Mann-Whitney Test	The Median test	3	Uniform
Mann-Whitney Test	The Median test	0.75	Laplace

Table: A.R.E of Test 1 to Test 2

Hypothesis Testing: Properties

Test 1	Test 2	A.R.E.	Assumptions
Sign Test	Paired t test	0.637	diff. are i.i.d normal
Sign Test	Wilcoxon signed test	0.667	diff. are i.i.d normal
Sign Test	Paired t test or Wilcoxon signed	0.333	diff. are i.i.d uniform (light-t
Sign Test	Paired t test	2	diff. are i.i.d Laplace (heavy
Sign Test	Wilcoxon signed	1.333	diff. are i.i.d Laplace (heavy
Cox-Stuart Trend	t test on regression coefficient	0.78	normal
Cox-Stuart Trend	Spearman's or Kendall's rank	0.79	normal
Cox-Stuart Trend	Spearman's or Kendall's rank	0.79	normal
The Median test	one-way ANOVA	0.64	Normal
The Median test	one-way ANOVA	2	Laplace (heavy-tailed)
Mann-Whitney Test	**Two-Sample t test**	**0.955**	**Normal**
Mann-Whitney Test	Two-Sample t test	1	Uniform
Mann-Whitney Test	Two-Sample t test	1.5	Laplace
Mann-Whitney Test	The Median test	1.5	Normal
Mann-Whitney Test	The Median test	3	Uniform
Mann-Whitney Test	The Median test	0.75	Laplace

Table: A.R.E of Test 1 to Test 2

Tests based on the Binomial

Tests based on the Binomial

Binomial experiment

- n independent trials
- Each trial has two possible outcomes
- p is the probability of having a “success" for each trail.
- The binomial describes the probability of obtaining k successes in the n trials

Remark

Many experimental situations in the applied science may be modeled as a Binomial experiment.

- Customers enter a store and decide to buy or not a product.
- Animals given a certain medicine and either they are cured or not cured.

Tests based on the Binomial: Estimation of p

Binomial Test

Data n independent trials. Each outcome is in either "class 1" or "class 2"

Assumptions Binomial experiment assumptions

Test statistic We are interested in the probability of the outcome "class 1". The **test statistic is the number of observations** in "class 1", denoted **T**.

Null distribution $T \sim Binom(n, p = p^*)$ under the null hypothesis. Use Binomial distribution if $n \leq 20$, otherwise use the normal approximation $x_q = np + z_q\sqrt{np(1-p)}$

Binomial Test: Hypotheses

Two-tailed test

$$H_0 : p = p^*$$
$$H_1 : p \neq p^*$$

The rejection region is defined by t_1 and t_2 as follows:

$$P(T \leq t_1) \approx \alpha/2 = \alpha_1 \text{ and } P(T \leq t_2) \approx 1 - \alpha/2 = 1 - \alpha_2$$

Decision Rule If ($T_{Observed} \leq t_1$ OR $T_{Obs} > t_2$), then REJECT H_0

P-value $2 \times \min\{P(T \leq T_{Observed}); P(T \geq T_{Observed})\}$

The actual $\alpha = \alpha_1 + \alpha_2$

Binomial Test: Hypotheses

Upper-tailed test

$$H_0 : p \leq p^*$$
$$H_1 : p > p^*$$

The rejection region is defined by t as follows:

$$P(T \leq t) \approx 1 - \alpha$$

Decision Rule If $T_{Obs} > t$ then REJECT H_0

P-value $P(T \geq T_{Obs})$

Binomial Test: Hypotheses

Lower-tailed test

$$H_0 : p \geq p^*$$
$$H_1 : p < p^*$$

The rejection region is defined by t as follows:

$$P(T \leq t) \approx \alpha$$

Decision Rule If $T_{Obs} \leq t$ then REJECT H_0

P-value $P(T \leq T_{Obs})$

Binomial Test: Normal approximation

The p-values using the normal approximation when $n > 20$ are given using:

$$P(T \leq T_{Obs}) \approx P\Big(Z \leq \frac{T_{Obs} - np^* + 0.5}{\sqrt{np^*(1-p^*)}}\Big)$$

and

$$P(T \geq T_{Obs}) \approx 1 - P\Big(Z \leq \frac{T_{Obs} - np^* - 0.5}{\sqrt{np^*(1-p^*)}}\Big)$$

which includes 0.5 as a *correction of continuity* that improves the normal approximation to the binomial.

R: The function **binom.test** can be used to perform the test in **R**, which provides also the confidence intervals.

Quantile Test

Data Let $X_1, X_2, X_3, \ldots, X_n$ be a random sample. The data consist of observations on the X_i.

Assumptions The X_i are a random sample. The measurement scale of the X_i is at least ordinal.

Test statistic We have two statistics:

T_1 The number of observations $\leq x^*$

T_2 The number of observations $< x^*$

IF $T_1 = T_2$ then none of the numbers in the data exactly equal to x^*

Null distribution $T_i \sim Binom(n, p = p^*), i = 1, 2$ under the null hypothesis. Use Binomial distribution if $n \leq 20$, otherwise use the normal approximation

$x_q = np + z_q\sqrt{np(1-p)}$

Quantile Test: Hypotheses

Two-tailed test

$$H_0 : p^{*th} \text{ population quantile is } x^*$$
$$[H_0 : P(X < x^*) \leq p^* \text{ and } P(X \leq x^*) \geq p^*]$$
$$H_1 : p^{*th} \text{ population quantile is not } x^*$$

The rejection region is defined by t_1 and t_2 as follows:

$$P(T \leq t_1) \approx \alpha/2 = \alpha_1 \text{ and } P(T \leq t_2) \approx 1 - \alpha/2 = 1 - \alpha_2$$

Decision Rule If ($T_{Observed} \leq t_1$ OR $T_{Obs} > t_2$), then REJECT H_0

P-value $2 \times \min\{P(T \leq T_{Observed}); P(T \geq T_{Observed})\}$

The actual $\alpha = \alpha_1 + \alpha_2$

Quantile Test: Hypotheses

Upper-tailed test - T_2

$$H_0 : p^{*th} \text{population quantile is at least as greater as } x^*$$
$$H_0 : \boldsymbol{P(X < x^*) \leq p^*}$$
$$H_1 : p^{*th} \text{population quantile is less than } x^*$$
$$H_1 : \boldsymbol{P(X < x^*) > p^*}$$

The rejection region is defined by t as follows:

$$P(T \leq t) \approx 1 - \alpha$$

Decision Rule If $T_{Obs} > t$ then REJECT H_0

P-value $P(T \geq T_{Obs})$

Quantile Test: Hypotheses

Lower-tailed test - T_1

$$H_0 : p^{*th} \text{population quantile is not greater as } x^*$$
$$H_0 : \boldsymbol{P(X \leq x^*) \geq p^*}$$
$$H_1 : p^{*th} \text{population quantile is greater than } x^*$$
$$H_1 : \boldsymbol{P(X \leq x^*) < p^*}$$

The rejection region is defined by t as follows:

$$P(T \leq t) \approx \alpha$$

Decision Rule If $T_{Obs} \leq t$ then REJECT H_0

P-value $P(T \leq T_{Obs})$

Tolerance Limits

Confidence limits are limits within which we expect a given population parameter, such as the mean, to lie. **Tolerance limits** are limits within which we expect a stated proportion of the population to lie with a certain probability. Formally:

$$P(X_{(r)} \leq \text{at least a fraction q of the population} \leq X_{(n+m-1)}) \geq 1-\alpha \quad (4)$$

For one-sided tolerance limits, let either r or m equals to zero, where $X_{(0)}$ and $X_{(n+1)}$ are considered to be $-\infty$ and $+\infty$.

Tolerance Limits: Sample size

The tolerance limits can be used to find:

A sample size n

- A sample size n needed to have at least q proportion of the population between the tolerance limits with $1-\alpha$ probability.

$$n \approx \frac{1}{4}\chi^2_{1-\alpha;2(r+m)}\frac{1+q}{1-q} + \frac{1}{2}(r+m-1)$$

where $\chi^2_{1-\alpha;2(r+m)}$ is the quantile of a chi-squared random variable.

You can use exact values from the tables, for the one-sided tolerance limits use Table A5. For the two-sided limits use Table A6.

The tolerance limits can be used to find:

The percent q

- The percent q of the population that is within the tolerance limits, given n, $1 - \alpha$, r, and m:

$$q = \frac{4n - 2(r + m - 1) - \chi^2_{1-\alpha;2(r+m)}}{4n - 2(r + m - 1) + \chi^2_{1-\alpha;2(r+m)}}$$

R: The function **nptol.int {tolerance}** can be used to find either the sample size or the percent q .

Tests based on the Binomial: paired data

Sign Test

The sign test is a binomial test with $p = 0.5$. It is useful for testing whether one random variable in a pair (X, Y) tends to be larger than the other random variable in the pair.

Data Let $X_1, X_2, X_3, \ldots, X_n$ be a random sample. The data consist of observations on the X_i.

Assumptions The X_i are a random sample. The measurement scale of the X_i is at least ordinal.

Example

- Consider a clinical investigation to assess the effectiveness of a new drug designed to reduce repetitive behavior, we can compare time before and after taking the new drug.
- This test can be compared to the parametric t-paired test.

Sign Test

Data The data consists **a bivariate random sample** (X_i, Y_i), where n' is the number of the pairs. There should be some natural basis for pairing the observations; otherwise the X's and Y's are independent, and the more powerful Mann-Whitney test is more appropriate. Within each pair (X_i, Y_i) a comparison is made and the pair is classified as:

"+" if $X_i < Y_i$
"-" if $X_i > Y_i$
"o" if $X_i = Y_i$ (tie)

Assumptions

1. The bivariate random variables (X_i, Y_i) are mutually independent .
2. The measurement scale is at least ordinal.

Sign Test

Test Statistic T = Total number of +'s

Null distribution is the Binomial distribution with $p = 0.5$ and **n=the number of non-tied pairs**. For n less than or equal to 20, otherwise the normal approximation.

Sign Test: Hypotheses

Two-tailed test

$$H_0 : P(+) = P(-)$$
$$H_1 : P(+) \neq P(-)$$

The null hypothesis is interpreted as X and Y have the same location parameter, it can be rewritten as follows:

$$H_0 : E(X_i) = E(Y_i)$$

$$H_1 : E(X_i) \neq E(Y_i)$$

The rejection region is given by t and $n - t$ where: $P(Y \leq t) \approx \alpha/2$

Decision IF ($T_{Obs} \leq t$ or $T_{Obs} \geq n - t$) REJECT H_0

P-value $2 \times \min(P(Y \leq T_{Obs}), P(Y \geq T_{Obs}))$

Sign Test: Hypotheses

Upper-tailed test

$$H_0 : P(+) \leq P(-)$$
$$H_1 : P(+) > P(-)$$

Similarly:

$$H_0 : E(X_i) \geq E(Y_i)$$
$$H_1 : E(X_i) < E(Y_i)$$

The rejection region is defined by $n - t$ as follows:

$$P(Y \leq t) \approx \alpha$$

Decision IF $T_{Obs} \geq n - t$ REJECT H_0

P-value $P(Y \geq T_{Obs})$

Sign Test: Hypotheses

Lower-tailed test

$$H_0 : P(+) \geq P(-)$$

$$H_1 : P(+) < P(-)$$

Similarly:

$$H_0 : E(X_i) \leq E(Y_i)$$

$$H_1 : E(X_i) > E(Y_i)$$

The rejection region is defined by t as follows:

$$P(Y \leq t) \approx \alpha$$

Decision IF $T_{Obs} \leq t$ REJECT H_0

P-value $P(Y \leq T_{Obs})$

Sign Test: Normal Approximation

These probabilities can be found using TABLE of the binomial distribution if $n \leq 20$, otherwise use the normal approximation

$$t_q = 1/2(n + z_q\sqrt{n})$$

The p-values with the normal approximation are:

$$P(Y \leq T_{Obs}) \approx P\left(Z \leq \frac{2T_{Obs} - n + 1}{\sqrt{n}}\right)$$

and

$$P(Y \geq T_{Obs}) \approx 1 - P\left(Z \leq \frac{2T_{Obs} - n - 1}{\sqrt{n}}\right)$$

Goodness-of-fit Tests

Goodness of fit test

A test for goodness of fit is used to test if a random sample (from some unknown distribution $F(x)$) does come from a specified distribution $F^*(x)$.

Tests

- Chi-squared test
- Kolmogorov-Smirnov Test
- Lilliefors Test is a Kolmogorov-Smirnov test that is based on the z-score.
- Shapiro-Wilk Test for Normality

Chi-squared test

The oldest and best-known goodness-of-fit test, first presented by Pearson (1900). The Chi-squared test data consists of a random sample $X_1, X_2, X_3, ..., X_n$. These observations are **grouped in C classes**. We have $1 \times C$ **contingency table**. The measurement scale is at least nominal.

Classes	1	2	...	C	Total
Observed Frequencies	O_1	O_2	...	O_C	N
Expected Frequencies	E_1	E_2	...	E_C	N

where $E_j = p_j^* N$; p_j^* is the probability of a random observation in class j, under the null hypothesis. $j = 1, 2, \ldots, C$

Chi-squared test

Hypothesis

$$H_0 : P(X \in Classj) = p_j^* \quad \text{for all } j = 1, 2, \ldots, C$$

$$H_1 : P(X \in Classj) \neq p_j^* \quad \text{for at least one class}$$

The Test statistic

$$T = \sum_{i=1}^{C} \frac{(O_i - E_i)^2}{E_i}$$

$T \sim_{H_0} \chi^2_{C-1-K}$; K is the number of estimated parameters. **This approximation requires that the number of E_j be large enough**. There are many rules that defines “large enough". Cochran(1952) proposed none of $E_j < 1$ and 80% are > 5. We will use this rule in this class. We will combine some cells if this rule is not satisfied.

Decision Rule Reject H_0 if $T > \chi^2_{1-\alpha;C-1-K}$

Kolmogorov-Smirnov Test

The data consist of a random sample $X_1, X_2, X_3, ..., X_n$.

Hypothesis

$$H_0 : F(x) = F^*(x)$$

$$H_1 : F(x) \neq F^*(x)$$

The Test statistic Let $S(x)$ be the empirical distribution function based on the data.

$$T = \sup_x \mid F^*(x) - S(x) \mid \quad (5)$$

which is read "T equals the supremum, over all x, of the absolute value of the difference $F^*(x) - S(x)$." The quantiles of T can be found in KS tables or using R.

Decision Rule Reject H_0 if $T_{observed} > T_{1-\alpha}$

Shapiro-Wilk Test for Normality

Data The data consist of a random sample $X_1, X_2, X_3, ..., X_n$.

Hypothesis

$$H_0 : F(x) \text{ is normal with unspecified mean and variance}$$
$$H_1 : F(x) \text{ is nonormal}$$

Test Statistic The order statistic is given as $X^{(1)}, X^{(2)}, X^{(3)}, ..., X^{(n)}$ from the smallest to the largest observation in the sample.

$$W = \frac{\left(\sum_{i=1}^{k} a_i (X^{(n-i+1)} - X^{(i)}) \right)^2}{\sum_{i=1}^{n} (X_i - \bar{X})} \quad (6)$$

The quantiles of T can be found in tables of the Test or using R.

Contingency Tables

Methods based on Ranks