Software Engineering Institute | Carnegie Mellon

# Basic Statistics Reference—1

## Guide to Visualizing Data

| Purpose | Graphical Method |
|---|---|
| See relationships in data | Scatter plot |
| See time relationships | Time series run chart |
| See variation of Y with 1 X | Box Plot chart |
| See variation of Y with 2+ X's | Multi-vari chart |
| Prioritize 2+ X's to focus on | Pareto chart |
| Check normality of data | Normal plot |
| Predict relationships in data | Regression predicted line |

## Measurement Error Guide for Continuous Data

| If the %GRR is… | Then measurement error is… |
|---|---|
| <10% | Acceptable |
| between 10% & 30% | Unacceptable for "critical" measurements (You should improve the measurement process.) |
| >30% | Unacceptable |

## Measurement Error Guide for Discrete (Attribute) Data

### When should each formal statistical approach be used?

**Attribute data is on Nominal scale**

⇩

**Fleiss's Kappa statistic**

.....

**Attribute data is on Ordinal scale (each item has at least 3 levels)**

⇩

**Kendall's coefficients**

### Interpreting results of Kappa's or Kendall's coefficients

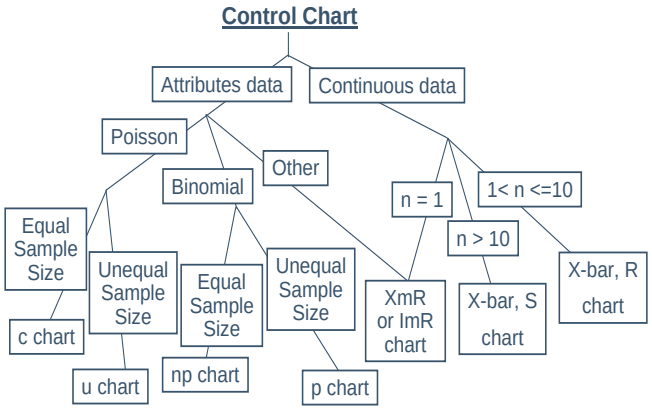
| | |
|---|---|
| When Result = 1.0 | perfect agreement |
| When Result > 0.9 | very low measurement error |
| When 0.70 < Result < 0.9 | marginal measurement error |
| When Result < 0.7 | too much measurement error |
| When Result = 0 | agreement only by chance |

### Interpreting the accompanying p value

**Null Hypothesis**: Consistency by chance; no association
**Alternative Hypothesis**: Significant consistency & association

*Thus, a p value < 0.05 indicates significant and believable consistency or association.*

## Statistical Process Control Chart




## Guide to Interpreting the p Value

| Method | Null | Alternative | P < 0.05 | P > 0.05 |
|---|---|---|---|---|
| Hypothesis Tests | No difference exists; no associations | Two items are different; association exists | Accept alternative | Accept null |
| Tests for Normality | Data follows normal distribution | Data does not follow normal distribution | Accept alternative | Accept null |
| ANOVA | No difference of Y across levels of x | Difference of Y exists between 1+ levels of x | Accept alternative | Accept null |
| Regression | x factor does not add value to model | X factor adds value to model | Accept alternative | Accept null |
| Chi-Square | Two discrete variables are not associated | Two discrete variables are associated | Accept alternative | Accept null |
| Logistic Regression | x factor does not add value; model has no significant x's | X factor adds value to model; model has 1+ significant x's | Accept alternative | Accept null |

# Basic Statistics Reference—2

## Types of Hypothesis Tests

| Data Type → | Interval or Ratio (Parametric Tests) | | Ordinal (Non-Parametric Tests) | | Nominal | Proportion |
|---|---|---|---|---|---|---|
| # Samples (Data groups) | Mean | Variance | Median | Variance / Fit | Similarity | Similarity |
| 1 Sample | 1-sample t test | 1-sample Chi-Square test | 1 sample Wilcoxon Signed Ranks test | Kolmogorov-Smirnov Goodness of Fit test | >2 cells: Chi-Square; =2 cells: Binomial Sign Test | 1 Proportions test |
| 2 Samples | Independent: 2-sample t test; Paired: Paired t test | Normal: F test; Not Normal: Levene test | Independent: Mann Whitney U test; Paired: Wilcoxon matched | = Medians: Siegel-Tukey test; ≠ Medians: Moses test | Fisher Exact test (1-way ANOVA); Chi-Square test | 2 Proportions test |
| 3+ Samples | ANOVA (1 & 2 way ANOVA; Balanced ANOVA; GLM) MANOVA (General & Balanced) | Normal: Bartlett test; Not Normal: Levene test | Independent: Kruskal-Wallis 1-way ANOVA; Paired: Friedman 2-way ANOVA | Van der Waerden Normal scores test | Chi-Square test | ANOM (Analysis of Means) |

## Hypothesis Tests *in Minitab*

| Data Type → | Interval or Ratio (Parametric Tests) | | Ordinal (Non-Parametric Tests) | | Nominal | Proportion |
|---|---|---|---|---|---|---|
| # Samples (Data groups) | Mean | Variance | Median | Variance / Fit | Similarity | Similarity |
| 1 Sample | Stat>Basic Statistics> 1 sample t | Stat>Basic Statistics>2 Variances> Summarized Data | Stat> Non-parametrics> 1-Sample Wilcoxon | Stat> ANOVA> Test for Equal Variances Or Stat> ANOVA> General Linear Model | Stat> Tables> Cross Tabulation and Chi-Square | Stat> Basic Statistics> 1 Proportion |
| 2 Samples | Independent: Stat>Basic Statistics>2 sample t; Paired: Stat>Basic Statistics> Paired t | Stat>Basic Statistics>2 Variances | Stat> Non-parametrics> Mann Whitney (Wilcoxon Matched Pair Test Not Automated) | Stat> ANOVA> Test for Equal Variances Or Stat> ANOVA> General Linear Model | Stat> Tables> Cross Tabulation and Chi-Square | Stat> Basic Statistics> 2 Proportions |
| 3+ Samples | Stat> ANOVA> *wide variety of choices* | Stat> ANOVA> Test for Equal Variances Or Stat> ANOVA> General Linear Model | Stat> Non-parametrics> Kruskal-Wallis or Friedman | Stat> ANOVA> Test for Equal Variances Or Stat> ANOVA> General Linear | Stat> Tables> Cross Tabulation and Chi-Square | Stat> ANOVA> Analysis of Means |

# Basic Statistics Reference—3

## Quantifying Relationships of X Factors with Y Outcomes

### ANOVA & MANOVA

| X (Discrete) | Y (Continuous): 1 Variable | Y (Continuous): >= 2 Variables |
|---|---|---|
| 1 Variable, 2 levels | t-Test | Hotelling's $T^2$, Discriminant Analysis |
| 1 Variable, >= 2 levels | 1-Way ANOVA | 1-Way MANOVA |
| >= 2 Variables | 2-Way ANOVA | 2-Way MANOVA |
| Mixture of Discrete & Continuous | ANCOVA | MANCOVA |

### ANOVA & MANOVA *in Minitab*

| X (Discrete) | Y (Continuous): 1 Variable | Y (Continuous): >= 2 Variables |
|---|---|---|
| 1 Variable, 2 levels | Stat>Basic Stats> sample t or paired t | Stat>Multivariate> Discriminant Analysis |
| 1 Variable, >= 2 levels | Stat>ANOVA>1 -way or 1-way (unstacked) | Stat>ANOVA>Balanced or General MANOVA |
| >= 2 Variables | Stat>ANOVA>2 -way | Stat>ANOVA>Balanced or General MANOVA |
| Mixture of Discrete & Continuous | | |

Begin here

| X | Y: Continuous | Y: Discrete |
|---|---|---|
| Discrete | ANOVA & MANOVA | Chi-Square & Logit |
| Continuous | Correlation & Regression | Logistic Regression |

### Correlation & Regression

| X (Continuous) | Y (Continuous): 1 Variable | Y (Continuous): >= 2 Variables |
|---|---|---|
| 1 Variable | Bivariate regression | |
| >= 2 Variables | Multiple regression | Multi-variable regression; Canonical analysis (CANONA) |

### Correlation & Regression *in Minitab*

| X (Continuous) | Y (Continuous): 1 Variable | Y (Continuous): >= 2 Variables |
|---|---|---|
| 1 Variable | Stat>Regression> Regression or Stepwise Regression | |
| >= 2 Variables | Stat>Regression> Regression or Stepwise Regression | |

# Basic Statistics Reference—4

## Data Types

Smaller sample sizes needed

| Type of Data | | Description | Examples |
|---|---|---|---|
| **Attribute or Categorical Data** | **Nominal** | Categories or buckets of data with **no ordering** | defect types<br>language types<br>customers<br>document types |
| | **Ordinal** | Categories or buckets of data with **ordering** | severity ratings<br>priority ratings<br>customer satisfaction ratings<br>high, medium, or low ratings |
| **Continuous Data** | **Interval** | Data measured on a scale that has **equal intervals** | productivity<br>defect density<br>preparation rate<br>cycle time<br>size<br>test hours |
| | **Ratio** | Interval data that has an **absolute zero** | |

## Process Quick Reference

### Hypothesis Test Steps

1. Identify a null and alternative hypothesis, and data sample(s).
2. If **continuous data**, test each sample for normality, and test both samples for equal variance.
   a) **If not normal but equal variances**, then either: 1) transform data using Box-Cox transformation and conduct parametric hypothesis tests (remember to un-transform results to interpret hypothesis), or 2) conduct nonparametric hypothesis tests,
   b) **If not equal variances**, then conduct nonparametric hypothesis tests,
   c) **If normal and equal variances**, conduct parametric hypothesis tests.
3. If **discrete data**, conduct nonparametric hypothesis tests
4. Use ***"Hypothesis Decision Matrix"*** to decide which parametric or nonparametric test to conduct, and use the ***"p Value Summary chart"*** to interpret the result of test.

### Quantifying Relationships

1. Identify a sample or population of data to analyze for correlation or predictive relationships.
2. Use **quantifying relationships decision matrix** to determine which method to use.
3. Use **p value summary chart** to conclude result.

### Regression Analysis Steps

1. Identify a Y outcome variable and 1+ explanatory x factors to conduct a regression analysis.
2. Conduct the regression analysis.
3. Use **p value summary chart** to conclude which x factors add value to the model.
4. Review **adjusted R squared value** to see how much (%) of the Y behavior is explained by the model.
5. Review the **4-in-one residuals chart** to ensure residuals look okay.
6. If **residuals** do not look okay, investigate nonlinearity, discard model or proceed with caution regarding significant conclusions.

### ANOVA Steps

1. Identify a Y outcome variable and 1+ explanatory x factors to conduct ANOVA.
2. Conduct the ANOVA.
3. Use **p value summary chart** to conclude whether the Y is different by levels of the x factor.
4. Review **adjusted R squared value** to see how much (%) of the Y behavior is explained by the ANOVA model.
5. Review the **confidence intervals** of the various levels of the x factor to see which are common (i.e., overlap) vs. which are different (i.e., do not overlap).
6. Review the **residuals chart** to ensure residuals look okay.
7. If **residuals** do not look okay, investigate **nonlinearity,** discard model, or proceed with caution regarding significant conclusions.