Statistics and Probability

Nishanth Nair

December,2016

Overview

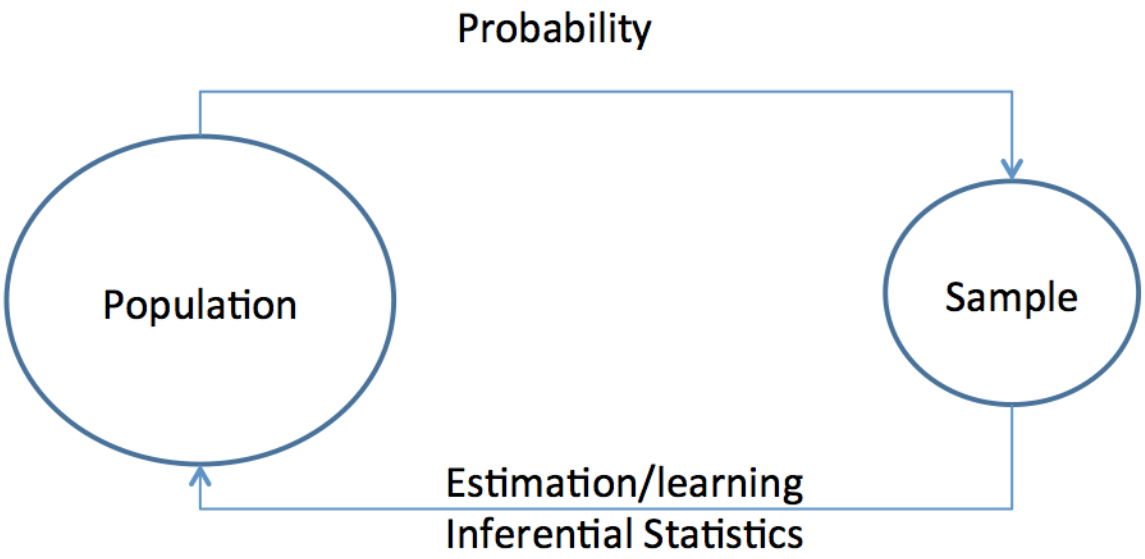

Figure 1: Relation between probability and Inferential Statistics

- Stem and Leaf Displays
  - Identification of a typical or representative value
  - Extent of spread about the typical value
  - Presence of gaps in data
  - Outliers
  - Number and location of peaks
  - Extent of symmetry in distribution of values
- Dotplots
  - Gives information about location,spread,extremes and gaps
- Histograms
  - Relative Frequency $= \frac{\text{Number of times the value occurs}}{\text{Number of observations in the data set}}$
  - Area of each rectangle is proportional to the relative frequency of the value.Total area of all rectangles is 1.
  - Number of classes $\approx \sqrt{\text{Number of Observations}}$

- Shapes: Unimodal, Bimodal, Multimodal
- Distribution: Symmetric, Positive skew, negative skew

Measures of Location:

1. Mean
   - $\bar{x} = \frac{x_1+x_2+...+x_n}{n}$
   - Extremely sensitive to outliers
   - Sum of deviations = $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$
2. Median
   - $\tilde{x}$ = Single middle value if n is odd | Average of middle values if n is even
   - Not sensitive to outliers
3. Quartiles
4. Percentiles
5. Trimmed Means

Measures of Variability:

1. Variance
   - Sample variance, $s^2 = \frac{\sum(x_i-\bar{x})^2}{n-1}$
   - Sample standard deviation, s = $\sqrt{s^2}$
   - Population variance, $\sigma^2 = \frac{\sum(x_i-\mu)^2}{N}$ (Note: N not N-1)
   - Population standard deviation, $\sigma = \sqrt{\sigma^2}$

Probability

We can associate with each event A, a probability P(A) measuring its relative likelihood. This probability should obey to the three following axioms:

1. For any event A, P(A) ≥ 0
2. If S denotes the sample space of an experiment then P(S) = 1
3. If events A and B are disjoint i.e. A$\cap$B=$\varnothing$, P(A$\cup$B) = P(A) + P(B)

- P($\varnothing$) = 0

- P(A∪B) = P(A) + P(B) - $P(A \cap B)$
- P(A) + P(A‘) = 1
- $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$
- Product Rule: If the first element of an ordered pair can be selected in n_1 ways and from each of these n_1 ways the second element in the pair can be selected in n_2 ways, then the total number of pairs is $n_1 n_2$
- An ordered subset is called permutation.An unordered subset is a combination.
- $P_{k,n} = \frac{n!}{(n-k)!}$
- $C_{k,n} = \binom{n}{k} = \frac{P_{k,n}}{k!} = \frac{n!}{k!(n-k)!}$
- $\binom{n}{n} = \binom{n}{0} = 1$

Conditional Probability

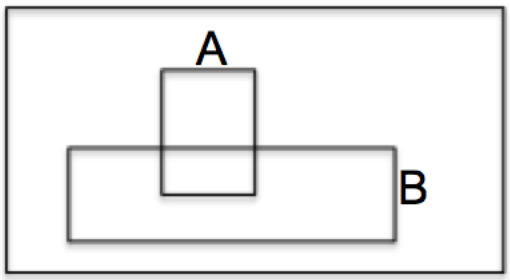

Figure 2: Conditional Probability

- $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- P(A‘|B) = 1 - P(A|B)
- A causal connection between A and B means: $P(A|B) \geq P(A)$
- If A and B are independent:
  - $P(A|B) = P(A)$
  - P(A ∩ B) = P(A).P(B) (Applies to multiple events)
- Law of Total Probability: If $A_1, ...A_k$ are mutually exclusive and exhaustive events, then for any event B,

$$P(B) = \sum_{i=1}^{k} P(B|A_i).P(A_i)$$

- Bayes Theorem: If $A_1, ...A_k$ are mutually exclusive and exhaustive events, given that B has occurred

$$P(A_j|B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(B|A_j).P(A_j)}{\sum_{i=1}^{k} P(B|A_i).P(A_i)}, j = 1...k$$

Discrete Random Variables

- Given a discrete random variable (rv) X which takes on values in S = $x_1, ...x_n$, its probability mass function is defined by:

$$p(x_i) = P(X = x_i)$$

- The cumulative distribution function, F(x) is the probability that the observed value of X is atmost x:

$$F(x) = P(X \leq x) = \sum_{y:y \leq x} p(y)$$

- $P(a \leq X \leq b) = F(b) - F(a) - 1$
- The expected value or mean (Weighted Average) of X, denoted by E(X) or μ_X is:

$$E(X) = \sum_{x \in D} x.p(x)$$

- Expected value of a function of X, h(X) is given by:

$$E[h(X)] = \sum_{D} h(x).p(x)$$

- E(aX+b) = a.E(X) + b
- Variance of X, denoted by V(X) or σ_X^2 is

$$V(X) = \sum_{D}(x - \mu)^2.p(x) = E[(x - \mu)^2] = E(X^2) - [E(X)]^2$$

- $V(aX + b) = a^2.V(X) \implies \sigma_{aX+b} = |a|.\sigma_X$
- V(constant) = 0

Continuous Random Variables

- If X is a continuous rv in R, the probability distribution or probability density function is given by

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx$$

- f(x) is the density curve or density function
  - $f(x) \geq 0 \,\forall x$
  - $\int_{-\infty}^{\infty} f(x)\,dx = 1$
- The cumulative distribution function, F(x) is the area under the density curve to the left of x:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)\,dx$$

- P(X > a) = 1 - F(a)
- P(X = a) = 0
- $P(a \leq X \leq b) = F(b) - F(a)$
- $F`(x) = f(x)$, if $F`(x)$ exists
- If p ∈ [0,1], the (100p)th percentile, denoted by $\eta(p)$ is

$$p = F[\eta(p)] = \int_{-\infty}^{\eta(p)} f(x)\,dx$$

- Median, $\tilde{x}$ is 50th percentile i.e half the area under the density curve is to the left of $\tilde{x}$ and other half to the right of $\tilde{x}$.
- The expected value or mean of X, denoted by E(X) or μ_X is:

$$E(X) = \int_{-\infty}^{\infty} x.f(x)\,dx$$

- Expected value of a function of X, h(X) is given by:

$$E[h(X)] = \int_{-\infty}^{\infty} h(x).f(x)\,dx$$

- Variance of X, denoted by V(X) or σ_X^2 is

$$V(X) = \int_{-\infty}^{\infty} (x-\mu)^2.f(x)\,dx = E[(x-\mu)^2] = E(X^2) - [E(X)]^2$$

Some discrete distributions

1. Bernoulli Distribution
2. Binomial Distribution
3. Hypergeometric Distribution
4. Poisson Distribution

Some continuous distributions

1. The Uniform Distribution
2. Exponential Distribution
3. Normal Distribution
4. Standard Normal Distribution
5. Gamma Distribution
6. Weibull Distribution
7. Lognormal Distribution
8. Beta Distribution
9. Chi Squared

Joint Probability Distributions

- If $X_1, X_2...X_n$ are discrete random variables,the joint pmf is given by

$$p(x_1, x_2..x_n) = P(X_1 = x_1 \cap X_2 = x_2.... \cap X_n = x_n)$$

- $p(x_1, x_2..x_n) \geq 0$
- $\sum_{X_1} ... \sum_{X_n} p(x_1, x_2..x_n) = 1$
- If A is a set consisting of pairs of (x,y), then

$$P[(X, Y) \in A] = \sum \sum_{(x,y)\in A} p(x, y)$$

- If $X_1, X_2...X_n$ are continuous random variables,the joint pdf is given by

$$P(a_1 \leq X_1 \leq b_1, ..., a_n \leq X_n \leq b_n) = \int_{a_1}^{b_1} .... \int_{a_n}^{b_n} f(x_1, .., x_n)\, dx_1..dx_n$$

Conditional Distributions

- Let X and Y be discrete random variables with pmf p(x,y)
  - The marginal pmf of X and Y are given by

$$\begin{aligned} p_X(x) &= \sum_y p(x, y) \\ &= \sum_y p_Y(y).p(x|y) \end{aligned}$$

$$\begin{aligned} p_Y(y) &= \sum_x p(x, y) \\ &= \sum_x p_X(x).p(y|x) \end{aligned}$$

  - The conditional pmf of Y given that X=x is

$$\begin{aligned} p_{Y|X}(y|x) &= \frac{p(x, y)}{p_X(x)} \\ &= \frac{p_Y(y).p_{X|Y}(x|y)}{p_X(x)} \end{aligned}$$

  - If X and Y are independent,

$$p(x, y) = p_X(x).p_Y(y)$$

- Let X and Y be continuous rv's with joint pdf f(x,y),
  - The marginal pdf of X and Y are given by

$$f_X(x) = \int_y f(x,y)\,dy \;\; -\infty < x < \infty$$

$$f_Y(y) = \int_x f(x,y)\,dx \;\; -\infty < y < \infty$$

  - The conditional pdf of Y given that X=x is

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} \;\; -\infty < y < \infty$$

  - The above also gives

$$f(x,y) = f(x).f(y|x) = f(y).f(x|y)$$

  - If X and Y are independent,

$$f(x,y) = f_X(x).f_Y(y)$$

- Chain Rule of probability

$$\begin{aligned} p(x,y,z) &= p(x).p(y|x).p(z|x,y) \\ &= p(z).p(y|z).p(x|y,z) \\ &= ... \end{aligned}$$

- Alternatives using the above

$$p(x) = \sum_{y,z} p(x,y,z) = \sum_{y,z} p(y,z).p(x|y,z) = ..$$

$$p(x|y,z) = \frac{p(x|z).p(y|x,z)}{p(y|z)}$$

- If X and Y are independent and event Z is given, the following hold
  - p(x|y,z) = p(x|z)
  - p(y|x,z) = p(y|z)
  - p(x,y|z) = p(x|z).p(y|z)
- Expected value of h(X,Y) is

$$\begin{aligned} E[h(X,Y)] &= \sum_x \sum_y h(x,y)p(x,y) \text{ (Discrete case)} \\ &= \int_x \int_y h(x,y)f(x,y)\,dx\,dy \text{ (Continuous case)} \end{aligned}$$

- Covariance of X and Y(Expected product of deviations)

$$\begin{aligned} cov(X,Y) &= E[(X-\mu_x)(Y-\mu_y)] \\ &= \sum_x \sum_y (x-\mu_x).(y-\mu_y)p(x,y) \text{ (Discrete case)} \\ &= \int_x \int_y (x-\mu_x).(y-\mu_y)f(x,y)\,dx\,dy \text{ (Continuous case)} \end{aligned}$$

  - $cov(X,X) = E[(X-\mu)^2] = V(X)$
  - $cov(X,Y) = E(XY) - \mu_x.\mu_y$
  - Correlation coefficient of X and Y

$$corr(X,Y) = \rho = \frac{cov(X,Y)}{\sigma_x.\sigma_y}$$

  - $-1 \leq \rho \leq 1$, $\rho = 0 \implies$ X and Y are uncorrelated
  - corr(aX+b,cY+d) = corr(X,Y)
  - if X and Y are independent, $\rho = 0$. Note: $\rho = 0$ does not imply independence.
  - $\rho = 1$ or -1 iff Y=aX+b, a $\neq 0$

- Conditional Expectation
  - The conditional expectation of X given Y=y is

$$\begin{aligned} E(X|Y) &= \sum_x x.p(x|y) \text{ (Discrete case)} \\ &= \int_x x.f(x|y)\,dx \text{ (Continuous case)} \end{aligned}$$

  - Tower property of conditional expectations(Law of iterated expectations)

$$\begin{aligned} E(X) &= E[E(X|Y)] \\ &= \sum_y E(X|Y).p(x|y) \text{ (Discrete case)} \\ &= \int_y E(X|Y).f(x|y)\,dy \text{ (Continuous case)} \end{aligned}$$

  - E(XY|X) = X.E(Y|X)
  - E(XY) = E[E(XY|X)] = E[X.E(Y|X)]

Sampling Distribution

- X_i's are said to be Idependent and Identically Distributed (iid) or form a random sample if:
  - X_i's are independent
  - Every X_i has the same probability distribution
- Notation and definitions:
  - $X_1, ...X_n$ form a random sample from a population.(are iid)
  - Sample Mean, $\bar{x}$
  - Sample variance, s^2
  - Sample standard deviation, s
  - Sample size, n
  - Population Mean, μ
  - Population Variance, σ^2
  - Population Standard Deviation, σ
  - Number of samples, N
  - Sampling distribution of $\bar{X}$:
    - $E(\bar{x}) = \mu_{\bar{x}} = \sum \bar{x}.p(\bar{x}) = \mu$
    - v($\bar{x}$) =$s^2 = \sum(\bar{x} - \mu_{\bar{x}})^2.p(\bar{x}) = \sum(\bar{x}^2.p(\bar{x})) - \mu_{\bar{x}}^2 = \sigma^2/n$
    - E(s^2) = $\mu_{s^2} = \sigma^2$
    - Standard error of the Mean, s = $\sigma_{\bar{x}} = \sigma/\sqrt{n}$
    - If $T_o = X_1 + X_2 + ... + X_n$
      - E(T_o) = n.μ
      - V(T_o) = n.σ^2
      - $\sigma_{T_o} = \sqrt{n}.\sigma$
- Central Limit Theorem: If $X_1, ...X_n$ form a random sample from a population with mean μ and variance σ^2. If n is sufficiently large (n > 30), $\bar{X}$ has a normal distribution with $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. T_o also has a normal distribution with μ_{T_o} = n.μ and $\sigma_{T_o} = \sqrt{n}.\sigma$ (Note: Assuming there are a large number of samples like X)
- For a Binomial Random Variable (i.e a random sample from a Bernoulli Distribution), with probability of success p, CLT applies if np $\geq$ 10 and n(1-p) $\geq$ 10.

- For a poisson rv (sum of independant Poisson rv's has a poisson distribution), if μ is sufficiently large (> 20), then the distribution is approximately normal and CLT applies.
- If X_i are independent random variables,
  - $s^2 = \sum (X_i - \bar{X})^2$
  - $p(\bar{X} = \mu) = \sum p(\bar{X}_i = \mu)$
  - $p(S^2 = s^2) = \sum p(\bar{S_i^2} = s^2)$

Distribution of a linear combination

- If $X_1, X_2 ... X_n$ are n random variables, Y = $a_1X_1 + a_2X_2 + ... a_nX_n$ is a linear combination of X_i's
  - If X_i's are independent or not
    - E$(a_1X_1 + a_2X_2 + ... a_nX_n) = a_1E(X_1) + a_2E(X_2) + ... a_nE(X_n)$
  - If X_i's are independent
    - V$(a_1X_1 + a_2X_2 + ... a_nX_n) = a_1^2V(X_1) + a_2^2V(X_2) + ... a_n^2V(X_n)$
  - For any X_i
    - V$(a_1X_1 + a_2X_2 + ... a_nX_n) = \sum_i \sum_j a_i a_j cov(X_i, X_j)$
- $E(X_1 - X_2) = E(X_1) - E(X_2)$
- $V(X_1 - X_2) = V(X_1) + V(X_2)$
- If Y = aX+b,

$$cor(X, Y) = \begin{cases} 1 & a > 0 \\ -1 & a < 0 \end{cases}$$

Point Estimations

- Given a statistical parameter θ, the value obtained for θ from a given sample is called the point estimate of θ. The selected statistic is called the point estimator of θ, represented as $\hat{\theta}$
  - For example, $\hat{\mu} = \bar{X}$ is the estimator for the sample mean
- $\hat{\theta}$ is a function of the sample, so its a random variable.

$$\hat{\theta} = \theta + \text{error in estimation}$$

- Ideally, we would want to find an estimator with the least error(Mean squared error or MSE).But since this is not possible in many cases, we look for an unbiased estimator.
- $\hat{\theta}$ is an unbiased estimator if $\mathrm{E}(\hat{\theta}) = \theta$
  - $\mathrm{E}(\hat{\theta})$ - θ is the bias in the estimator or error in estimation
- When X is a binomial rv with parameters n and p, the estimator $\hat{p} = X/n$ is an unbiased estimator of p.
- When X_i is a random sample from a distribution with mean μ and variance σ^2, the estimator

$$\hat{\sigma^2} = \frac{\sum(X_i - \bar{X})^2}{n-1}$$

  is an unbiased estimator for σ^2
- Choosing an unbiased Estimator
  - Choose the estimator with the minimum variance.This is the Minimum Variance Unbiased Estimator(MVUE).
  - When X_i is a random sample from a Normal Distribution, $\hat{\mu} = \bar{X}$ is the MUVE for μ
  - When X_i is a random sample from a Cauchy Distribution, $\hat{\mu} = \tilde{X}(Median)$ is the best estimator for μ. MUVE is unknown.
  - When X_i is a random sample from a uniform Distribution, $\hat{\mu} = \bar{X_e}$(Average of extremes) is the best estimator for μ.
  - A trimmed mean with small trimmed percentages is a robust estimator for μ.Its not the best but works well for most distributions
  - Bootstrap Estimate
  - Standard Error

Methods to obtain point estimators

1. Method of Moments:

In this method, equate certain sample characteristics to the corresponding population expected values.Solve the equations for the unknown parameter values to obtain the estimators.

- Let $X_1, ...X_n$ be a random sample from a distribution with pdf f(x).The kth population moment or kth moment of the distribution of f(x) is E(X^k). The kth sample moment is $(1/n)\sum_{i=1}^{n} X_i^k$
  - First population moment is E(X) = μ
  - First sample moment is $\sum X_i/n = \bar{X}$
  - Second population moment is E(X^2)
  - Second sample moment is $\sum X_i^2/n$
- First write down the population moments in terms of the unknown parametrs

$$E[X^k] = f_k(\hat{\theta_1}, \hat{\theta_2}....\hat{\theta_d}) \ , k = 1..d$$

  The method of moments estimators $\hat{\theta_1}, \hat{\theta_2}....\hat{\theta_d}$ are then obtained by equating the population moments to the sample moments

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k = f_k(\hat{\theta_1}, \hat{\theta_2}....\hat{\theta_d}) \ , k = 1..d$$

2. Maximum Likelihood Estimator

Let $X_1, ...X_n$ be a random sample from a distribution with pdf

$$f(x_1, x_2..., x_n; \theta_1, \theta_2...\theta_n)$$

Where $\theta_1, \theta_2...\theta_n$ have unknown values.

When x_i's have known sample values, the pdf becomes a function of θ and it is called the likelihood function.

The maximum likelihood estimates(MLEs), $\hat{\theta_1}, \hat{\theta_2}....\hat{\theta_n}$ are those values of the parameter that maximize the likelihood function.So

$$f(x_1, x_2..., x_n\ \hat{\theta_1}, \hat{\theta_2}....\hat{\theta_n}) \leq f(x_1, x_2..., x_n\ \theta_1, \theta_2...\theta_n)$$

- Its not necessary MLE yields an unbiased estimator.
- It is the fact that the MLE has the fastest possible rate of convergence among all possible estimators of an unknown statistical parameter that has let to its adoption as the gold standard estimator.

- When the sample size is large, the maximum likelihood estimator is atleast approximately unbiased i.e. E($\hat{\theta}$) $\approx \theta$ and has variance that is exactly or atleast approximately the MVUE of θ

Confidence Intervals based on a single sample

- General form: CI = sample estimate ± margin of error = sample estimate ± (multiplier) (standard error)
- 100(1-α)% CI
  - for mean, μ of a normal distribution with known σ

$$\bar{x} \pm z_{\alpha/2}.\sigma/\sqrt{n}$$

    * CI width, w = $2.z_{\alpha/2}.\sigma/\sqrt{n}$
    * The half width ($z_{\alpha/2}.\sigma/\sqrt{n}$) is called the bound on the error of estimation,B.That is, the estimate will be no farther than B from the population mean,μ
    * Sample size required to have a CI width w = $(2.z_{\alpha/2}.\sigma/w)^2$
    * Sample size required to have a bound B = $(z_{\alpha/2}.\sigma/B)^2$
  - If n is large (> 40), by CLT $\bar{X}$ has an approximately normal distribution.In this case,CI is

$$\bar{x} \pm z_{\alpha/2}.s/\sqrt{n}$$

  - Generally, with a t-distribution, CI is

$$\bar{x} \pm t_{\alpha/2,n-1}.s/\sqrt{n}$$

    * $t_{\alpha,v}$ is the t-critical value with v degrees of freedom and α is the measurement on the x-axis. This is the area under the t-curve to the right of α

Tests of hypotheses based on a single sample

- Statistical significance: Sample statistics vary from the specificed population parameters to the extent that it is unlikely that the results obtained were due to random sampling error, rather we conclude that the differences observed in the sample were due to actual differences in the population
- p-value: Probability that the observed sample statistic (or a statistic more extreme) was randomly obtained from a population with the hypothesized parameter.A p-value can be thought of as the following conditional probability: P(Rejecting $H_o|H_o$ is true)
- A sample result is called statistically significant when the p-value for a test statistic is less than or equal to the level of significance, this is also known as the alpha (α) level. Typically $\alpha = 0.05$
- Hypotheses testing procedure
  - Check any necessary assumptions and write null and alternative hypotheses
  - Calculate an appropriate test statistic (typically the difference observed in the sample divided by a standard error)
  - Determine a p-value associated with the test statistic
  - Decide to reject or fail to reject the null hypothesis: If p $< \alpha$ reject the null hypothesis. If p $> \alpha$ fail to reject the null hypothesis.
  - State a "real world" conclusion

Hypotheses testing for Mean

- STEP 1: Check any necessary assumptions and write null and alternative hypotheses

Assumption:
Data must be quantitative and randomly sampled from a population that is approximately normally distributed.
Assume significance level of α for the test

Research Question	Mean diff from μ_o	Mean $> \mu_o$	Mean $< \mu_o$
H_o	$\mu = \mu_o$	$\mu = \mu_o$	$\mu = \mu_o$
H_a	$\mu = \mu_o$	$\mu > \mu_o$	$\mu < \mu_o$
Test Type, Direction	Two Tailed, Non Directional	Right Tailed, Directional	Left Tailed, directional

- STEP 2: Calculate an appropriate test statistic

Use the one sample t-test

$$t = \frac{\bar{x} - \mu_o}{s/\sqrt{n}}$$

- STEP 3:Determine a p-value associated with the test statistic

Using a t-test, p-value is the area under the t_{n-1} curve.

p-value is determined as follows in R:

For a two tailed test:

```
T <- -1.198 # assume calculated in step 2
n <- 57 # Assume sample size

p_value <- 2*(pt(-abs(T),df=n-1))
```

For a single tailed test:

```
# Left Tailed test (Less than)
p_value <- pt(abs(T),df=n-1)


# Right Tailed test(Greater than)
p_value <- pt(-abs(T),df=n-1)
```

- STEP 4: Decide between the null and alternative hypotheses

If p $\leq \alpha$ reject the null hypothesis.

If p $> \alpha$ do not reject the null hypothesis.

- STEP 5: Conclude

Note: When the null hypothesis cannot be rejected, there are two possible cases:

1. the null hypothesis is true
2. the sample size is not large enough reject the null hypothesis.

Errors in testing

- Type I error: rejecting H_o when H_o is really true
  - Denoted by α (i.e. level of significance)
  - α = p(Type I error)
- Type II error: Failing to reject H_o when H_o is really false
  - Denoted by β
  - β = p(Type II error)

Cautions About Significance Tests

- If a test fails to reject H_o , it does not mean that H_o is true – it just means we do not have compelling evidence to refute it. This is especially true for small sample sizes.
- Our methods depend on a normal approximation. If the underlying distribution is not normal (e.g., heavily skewed, several outliers) and our sample size is not large enough to offset these problems (recall the Central Limit Theorem) then our conclusions may be inaccurate.
- When comparing a p value to an alpha level we are examining statistical significance. If results are statistically significant, that means that it is unlikely that our sample came from a population with the hypothesized value. It tells us that our sample is likely different from the hypothesized population. However, it does not tell us how different.With a very large sample size, almost any difference will be statistically significant. Thus, it is important to pay attention to the size of the difference when using statistics to make decisions in real life.

Power

- Power is the probability of rejecting the null hypotheses, given that the null hypothesis is false; in other words, the probability of correctly rejecting H_o
- The power of a test can be increased in a number of ways,
  - increasing the sample size,
  - decreasing the standard error,
  - increasing the difference between the sample statistic and the hypothesized value, or
  - increasing the alpha level.
- Relation between α and β
  - If the sample size is fixed, then decreasing α will increase β.
  - If we want both α and β to decrease, then we must increase the sample size.
- Power = 1 - β
- Examples:
  - If the power of a statistical test is increased, for example by increasing the sample size, the probability of a Type II error decreases since β = 1 - power

Independent and Dependent Tests

Figure 3: Problems and Hypothesis

- Research Problem -> Questions -> hypothesis -> Variables
- Criteria for a research problem
  - What will we learn that we already do not know
  - Why is it worth knowing
  - How will we know the conclusions are valid
- Justifying a research problem
  - Explain what is not known about the problem
  - Why does the problem matter
  - Show this is actually a problem - statistics,literature etc?
- Defining Research Questions
  - An interrogative sentence or statement that asks: What relation exists between two or more concepts
- Hypothesis are used to relate specific variables
  - Hypothesis statements contain two or more variables that are measurable or potentially measurable and that specify how the variables are related.(Kerlinger,1986)
  - Deals with variables not concepts, measurable
- Propositions vs Hypothesis

– Propositions link concepts; Hypothesis link variables

Determining the test type

- If the measurements are **categorical** (e.g. Yes or No to the question "Do you smoke?") and taken from two **distinct groups** (e.g. Female, Male) the analysis will involve comparing two independent proportions.
- If the measurements are **quantitative** (e.g. GPA) and taken from two **distinct groups** (e.g. Graduate, Undergraduate) the analysis will involve comparing two independent means.
- If the measurements are **quantitative** (e.g. Weight) and taken **twice from each subject** (e.g. subject's weight before and after dieting) the analysis will involve comparing two dependent means.
- Finally, there are situations that involve **dependent proportions** (e.g. we ask subjects whether they support a particular candidate before and after a political debate and compare their responses across Sex).

1. Comparing population proportions: Independent Samples

- **Categorical Data**: Observations are classified into categories so that the data set consists of frequency counts for the categories.
- When we have a categorical variable of interest measured in two populations, it is quite often that we are interested in comparing the proportions of a certain category for the two populations.
- Point estimate is the difference between the two sample proportions is given by:

$$\hat{p_1} - \hat{p_2} = \frac{x_1}{n_1} - \frac{x_2}{n_2}$$

- The standard deviation is given by, $SE_(\hat{p_1} - \hat{p_2})$:

$$\sqrt{\frac{\hat{p_1}.(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}.(1-\hat{p_1})}{n_2}}$$

- **Assumption**: If $n_1.\hat{p_1}$,$n_1.(1-\hat{p_1})$,$n_2.\hat{p_2}$,$n_2.(1-\hat{p_2})$ are all atleast 10, the distribution of $\hat{p_1} - \hat{p_2}$ is approximately normal and we can use the z-test.
- Z-statistic is given by:

$$Z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{p^*.(1-p^*).(\frac{1}{n_1} + \frac{1}{n_2})}}$$

  where,

$$p^* = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1}{n_1 + n_2}.\hat{p_1} + \frac{n_2}{n_1 + n_2}.\hat{p_2}$$

- 100(1-α)% confidence interval of $p_1 - p_2$ is given by:

$$\hat{p_1} - \hat{p_2} \pm z_{\alpha/2}.SE_(\hat{p_1} - \hat{p_2})$$

  where,

$$SE_(\hat{p_1} - \hat{p_2}) = \sqrt{\frac{\hat{p_1}.(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}.(1-\hat{p_1})}{n_2}}$$

2. Comparing population Means: Independent samples

- Point estimate for $\mu_1 - \mu_2$ is $\bar{x_1} - \bar{x_2}$
- If the following assumptions hold:
  - Sample sizes are large or samples from each population are normal
  - Samples are taken independently
  - Population variances or standard deviations are known

Then, $\bar{y_1} - \bar{y_2}$ is normal with mean $\mu_1 - \mu_2$ and standard deviation

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

2 sample t-test using pooled variances

- Use the 2-sample pooled variance t-test as follows:
  - If variances of population 1 and population 2 are nearly equal, OR
  - Ratio of variances lies between 0.5 and 2
  - Let n_1 be the sample size of population 1 with sample standard deviation s_1
  - Let n_2 be the sample size of population 2 with sample standard deviation s_2
  - The common or pooled standard deviation is given by:

$$s_p = \sqrt{\frac{(n_1 - 1)^2.s_1^2 + (n_2 - 1)^2.s_2^2}{n_1 + n_2 - 1}}$$

  - Standard error is given by:

$$SE = s_p.\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

  - t-statistic is given by:

$$t = \frac{\bar{x_1} - \bar{x_2}}{SE}$$

  - Degrees of freedom $= n_1 + n_2 - 1$
- General steps to test hypothesis:
  - Validate assumptions
    - Are samples independent?
    - are sample sizes > 30? if not, do samples come from normal population distributions?
    - Are population variances equal or does s1/s2 lie between 0.5 and 2

    - If above hold, use pooled variance t-test
  - Step 1: Form hypothesis
  - Step 2: Specify significance level(α)
  - Step 3: Compute t-statistic
  - Step 4: Compute p value
  - Step 5: Decide whether to reject H_o
  - Step 6: Formulate conclusion

- General steps to estimate the difference between population means:
  - Validate assumptions:
    - Samples are independent
    - are sample sizes > 30? if not, do samples come from normal population distributions?
    - Are population variances equal or does s1/s2 lie between 0.5 and 2
    - If above hold, use pooled variance t-test
  - Step 1: Find $t_{\alpha/2}$ with D.F $= n_1 + n_2 - 1$
  - Step 2: End points of (1-α).100 % confidence interval is:

$$\bar{x_1} - \bar{x_2} \pm t_{\alpha/2}.SE$$

- **What to do if assumptions are not valid**:
  - if the assumption of independent samples is violated:
    - If the samples are not independent but paired, we can use the paired t-test.
  - if the sample sizes are not large **and** the populations are not normal
    - use a non-parametric method to compare two samples
  - if the assumption of equal variances is violated
    - use the separate variances 2-sample t-test

2 sample t-test using seperate variances

- t-statistic is given by:

$$t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- Degrees of freedom is given by:

$$df = \frac{(n_1 - 1).(n_2 - 1)}{(n_2 - 1).C^2 + (1 - C)^2(n_1 - 1)}$$

  where,

$$C = \frac{s_1^2/n_1}{s_1^2/n_1 + s_2^2/n_2}$$

  Alternatively, d.f = smaller of (n_1-1) and (n_2-1)

3. Comparing population Means: Paired samples

- Assumptions:
  - Samples are paired
  - The differences of the pairs follow a normal distribution or the number of pairs is large (note here that if the number of pairs is < 30, we need to check whether the differences are normal, **but we do not need to check for the normality of each population**)
- Hypothesis:
  - $H_o : \mu_d = \Delta_0$
  - $H_1 : \mu_d > \Delta_0$
  - $H_2 : \mu_d < \Delta_0$
  - $H_3 : \mu_d \neq \Delta_0$
- t-Statistic
  - Let D be the differences between pairs of data, then
    - $\bar{d}$ is the mean of the differences.
    - Number of pairs = n
    - Standard deviation of D = s_D
    - Degrees of freedom = n-1
    - standard deviation of the sample differences, $s_{\bar{d}} = s_D/\sqrt{n}$
  - Then,t-statistic is given by
$$t = \frac{\bar{d} - \Delta_0}{s_D/\sqrt{n}}$$
  - Paired t-interval -
$$\bar{d} \pm t_{\alpha/2}.s_D/\sqrt{n}$$

Practical significance of t-Tests

- Effect size allows us to compare the relative effect of one test to another.
- Effect size is a measure of practical significance, not statistical significance.
- Effect size estimates a population parameter, and it is NOT affected by the sample size.
- Higher effect sizes mean less overlap between the two distributions.
- Sometimes large effect sizes can be worth exploring even if they are not statistically significant.
- Cohen's d and effect size corelation are standardized measures of effect size.Use these when the units are not informative.
- The difference between means ia another way to measure effect size(non standardized approach).
- **Cohen's d Effect Size**
  - d = $\frac{m_1 - m_2}{s_{pooled}}$

- $s_{pooled} = \sqrt{\frac{(n_1-1)^2.s_1+(n_2-1)^2.s_2}{n_1+n_2-2}}$
- Interpreting the values of d:
  - abs value of d greater than 0.8 are large,
  - abs value of d values around 0.5 are medium, and
  - abs value of d values less than 0.2 are small

- **Effect Size corelation (r)**
  - r = $\frac{t}{\sqrt{t^2+df}}$
  - abs value of r - 0.10 - small
  - abs value of r - 0.30 - Medium
  - abs value of r - 0.50 - Large

4. Comparing population proportions: Paired Samples

5. Comparing population Variances

6. Non-Parametric tests - Distribution free tests

- Advantages of non-parametric methods:
  - We can use the methods in situations where the assumptions of distribution theory is not supported.
  - They require few assumptions about the underlying distribution.
  - In some cases, Nonparametric methods are easier to apply.
  - They are generally robust to outliers.
  - Often the only methods available when the data consist of ranks, ordinal, or categorical data.

Permutation Principle

2 sample Permutation Test

- Quick rule of thumb to help determine which statistic to calculate.
  - If the data are approximately Normal, then use the difference of two means.
  - If the data are symmetric but outliers are present, then use the difference of two trimmed means.
  - If the data are not symmetric, then use the difference of two medians.

Type of Design	Parametric Tests	Nonparametric Tests
Two independent samples	Independent samples *t* test	Wilcoxon rank-sum test (Mann-Whitney test)
Two dependent Samples	Dependent samples *t* test	Wilcoxon signed-rank test

Figure 4: Parametric and Nonparametric Tests for Comparing Only Two Groups

Wilcoxon Rank-Sum Test (Equivalent to Mann-Whitney Test)

- Data are ranked from lowest to highest across the groups.
- If the same score occurs more than once, then all scores of the same value receive the average of the potential ranks for those scores.
- After assigning final ranks, we add up all the final ranks for each of the two groups.
- Then we subtract the mean rank for a group of the same size as our groups. (Otherwise larger groups would always have larger values.)
- W = sum of ranks − mean rank
- Interpretation:
  - Just as with a t test, the default is a two-sided test:
    - **Null hypothesis**: There is no difference in ranks.
    - **Alternative hypothesis**: There is a difference in ranks.
    - You can also do a one-directional test (if you hypothesize that one particular group will have higher ranks than the other).
    - There are always two values for W (one for each group).
    - The **lowest score** for W is typically used as the test statistic.
    - For small sample sizes (N < 40), R calculates the p value with Monte Carlo methods (i.e., simulated data are used to estimate the statistic).
    - For larger samples, R calculates the p value with a normal approximation method (it only assumes that the sampling distribution of the W statistic is normal, not the data).
- Effect Size calculation
  - Effect size correlation: $r = Z/\sqrt{n_1 + n_2}$
    - To calculate the effect size, simply divide the z statistic by the square root of the total sample size.

Wilcoxon Signed-Rank Test

- nonparametric equivalent of the dependent t-test.
- Unlike the dependent t test, the Wilcoxon signed-rank test deals with sums of ranks rather than averages of scores.
- The calculation of ranks is similar to the rank-sum test, except the focus is on the difference between the first scores and the second scores.
- These differences can be positive or negative. (Scores of zero are excluded.)
- The scores are ranked by absolute value.
- After ranking, the positive and negative ranks are summed separately.
- Between the positive and the negative sum of ranks, the smaller value is used as the test statistic.
- Effect Size Calculation:
  - We calculate effect size correlation only after we conduct the statistical test.
  - Effect size correlation, r = z statistic divided by the square root of the sample size
  - In R, we can use the qnorm() function to find the z statistic from our p value.

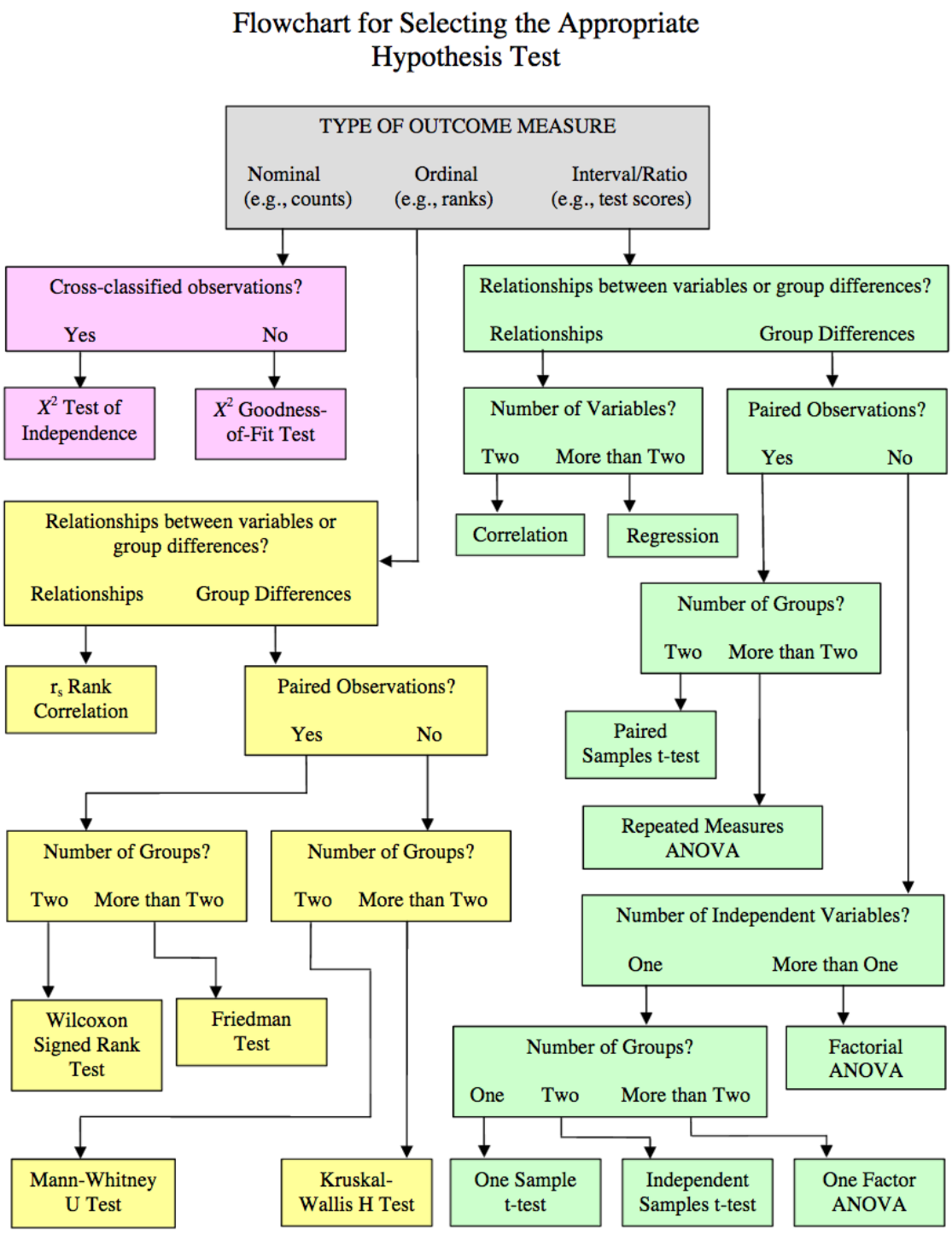

Figure 5: Selecting the appropriate hypothesis test

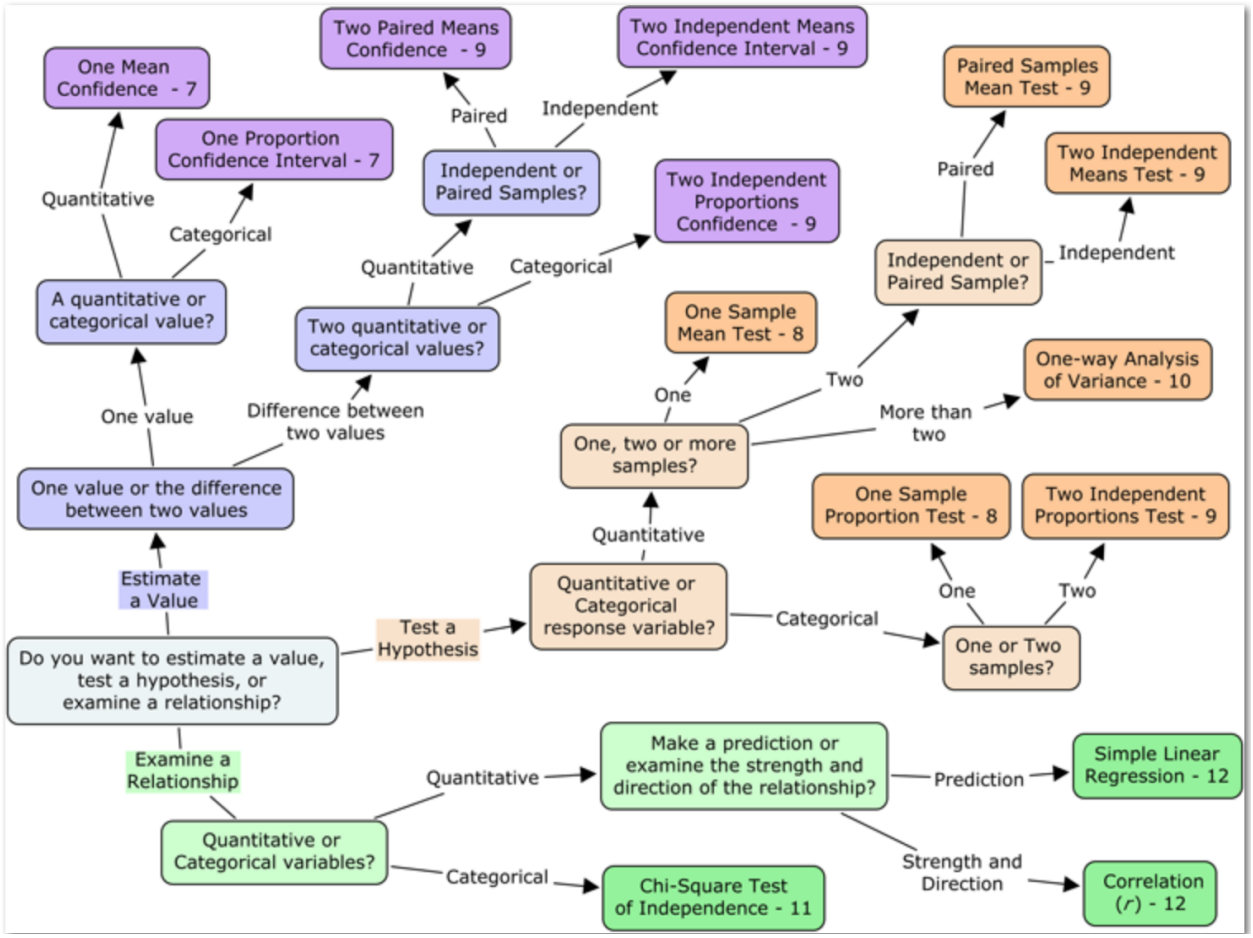

Figure 6: Determine appropriate statistical technique