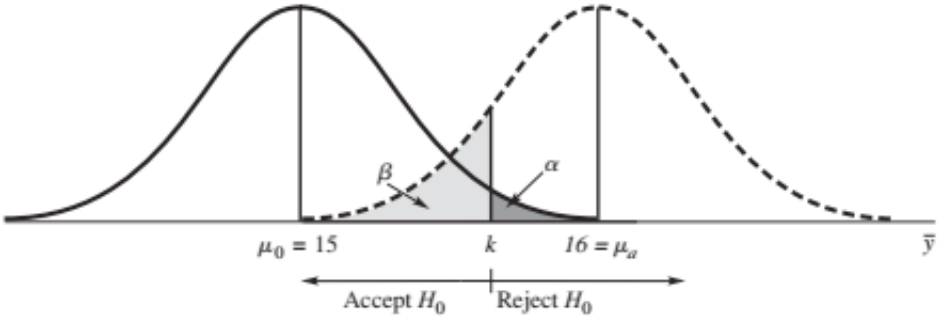
Chapter 9 - Testing Hypotheses and Goodness of Fit

Neyman-Pearson Paradigm

- Null hypothesis H_0 and alternative hypothesis H_A
  - **Type I error**: rejecting null when it is true, FP. The probability of type I error is the **significance level of the test** α - jury returns guilty when innocent. $\alpha = P(T(X_1, \ldots, X_n) > t_0 | H_0)$
  - **Type II error**: accepting the null hypothesis when false, FN. Denoted by β. Jury says innocent when guilty
    - $\beta = P(T(X_1, \ldots, X_n) \leq t_0 | H_A)$
    - Inherent trade off between Type I error and Type II - Type I maximized when Type II minimized and vice versa
  - p-value: $p = P(T(X_1, \ldots, X_n) \geq T(x_1, \ldots, x_n) | H_0)$ The chance of getting a value of T as large as the one we got or larger under the null hypothesis. Our test rejects when $p < \alpha$
  - **Power**: the probability that the null hypothesis is rejected when false, $1 - \beta$
  - Test statistic, statistic we are using as a test of the hypothesis. A function of sample we are going to test - eg. LRT
  - Null distribution: the probability distribution of the test statistic when the null hypothesis is true

1. Construct a test statistic T from our data, eg. for $H_0 : \mu = 0,\ H_1 : \mu \neq 0$, take $T = |\bar{X}|$
2. Construct probability of exceeding critical value, set an alpha for the test.
3. Either compute critical value from PDF of x and alpha (don't need x) or compute p-value from x and its PDF (don't need alpha)
4. Determine if we reject null or not

- A simple hypothesis: H_0 and H_1 each have parameter equal to a specific value, specify a complete probability distribution. If a hypothesis does not completely specify the probability distribution, the hypothesis is called a composite hypothesis. This means a hypothesis like this is Poisson distributed or not is composite, because the null hypothesis needs to be a specific distribution with specified parameter to be simple. It is convention to choose the simpler hypothesis to be the null.
- Neyman Pearson Lemma: Suppose that H_0 and H_A are simple hypotheses and that the test that rejects H_0 whenever the likelihood ratio is less than c and significance level α. Then any other test for which the significance level is less than or equal to α has power less than or equal to that of the likelihood ratio test.
- For simple hypothesis: We write down the likelihood ratio and observe that small values of it correspond in a one-to-one manner with extreme values of a test statistic, in this case X. Knowing the null distribution of the test statistic makes it possible to choose a critical level that produces a desired significance level α.
- p-value is the smallest alpha for which we reject the null hypothesis
  - critical value < test stat $\iff$ p-value < alpha
  - The p-value is the probability of a result as or more extreme than that actually observed if the null hypothesis were true.



- UMP: If the alternative H_1 is composite, a test that is most powerful for every simple alternative in H_1 is said to be uniformly most powerful.
  - In typical composite situations, there is no uniformly most powerful test.
  - Cannot by UMP against two sided alternative
- The alternatives $H_1 : \mu < \mu_0$ and $H_1 : \mu > \mu_0$ are called one-sided alternatives. The alternative $H_1 : \mu = \mu_0$ is a two-sided alternative.
- Confidence Intervals: μ_0 lies in the confidence interval for μ if and only if the hypothesis test accepts. In other words, the confidence interval consists precisely of all those values of μ_0 for which the null hypothesis $H_0 : \mu = \mu_0$ is accepted.
- To construct CI, want test statistic < CV instead of greater than since creating bounds of accepting null
- Hypothesis testing with samples:
  - $\Pr\left(\bar{X} - \frac{st_{(n-1)}^{1-\alpha/2}}{\sqrt{n}} \leqslant \mu \leqslant \bar{X} + \frac{st_{(n-1)}^{1-\alpha/2}}{\sqrt{n}}\right) = 1 - \alpha$
  - Since symmetric t distribution: $\Pr\left(\bar{X} + \frac{st_{(n-1)}^{\alpha/2}}{\sqrt{n}} \leqslant \mu \leqslant \bar{X} + \frac{st_{(n-1)}^{1-\alpha/2}}{\sqrt{n}}\right) = 1 - \alpha$
  - 2 Sample t-test estimating $\mu_x - \mu_y$ by $\bar{X} - \bar{Y}$.
    - $\bar{X} - \bar{Y} \sim N(\mu_x - \mu_y, \sigma^2(1/n + 1/m))$
    - $s_{pooled}^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}$
    - $\hat{\Delta} \pm t^{1-\alpha/2} s_{\text{pooled}} \sqrt{1/n + 1/m}$

Bayes Testing / LRT

- For simple hypotheses. Likelihood Ratio - For two distributions, the probability of observing an event from one over the other.
- $LR = \frac{P(\text{all our data}|H_0)}{P(\text{all our data}|H_1)}$

1. Use density functions of data in the LR
2. Plug in the parameters under H_0 and H_1, eg μ_1, μ_2
3. Simplify LR and determine how LR changes as a function of X
4. Assume X is distributed by the density given in H_0. We use this density in the test.
5. Set LRT greater than some constant critical value and determine CI for given alpha using some known distribution approximation (eg. normal, chi-square)

- This example is typical of the way that the Neyman-Pearson Lemma is used. We write down the likelihood ratio and observe that small values of it correspond in a one-to-one manner with extreme

values of a test statistic, fore example $\bar{X}$. Knowing the null distribution of the test statistic makes it possible to choose a critical level that produces a desired significance level α.
- The evidence provided by the data is contained in the likelihood ratio, which is multiplied by the ratio of prior probabilities to produce the ratio of posterior probabilities.

Generalized Likelihood Ratio Tests

- $\Lambda = \frac{\max_{\theta \in \omega_0}[\text{lik}(\theta)]}{\max_{\theta \in \Omega}[\text{lik}(\theta)]} \leq \lambda_0$ - Testing the parameter space for observations against some cutoff value. $\Lambda \leq 1$ since numerator is a subset of the denominator
- We have to work out the distribution of Λ under H_0 to get a significance level. This can be quite hard, so for iid data, we approximate using $-2log(\Lambda) \approx \chi^2_{(d)}$ under H_0 as $n \to \infty$.
    - DoF = dim Ω - dim ω_0 - if $\omega_0 \sim N(\mu, \sigma^2)$ specified then dim $\omega_0 = 0$ and dim Ω = number of free parameters
    - DoF d = # free parameters in H1 - # free parameters in H0. Example: N(2, 1) has 0 free parameters, $N(\sigma^2 + 1, \sigma^2)$ has 1, and $N(\mu, \sigma^2)$ has 2.
    - Intuition: the more free parameters you allow, the more the alternative can fit to the data and explain it. Therefore $-2log(\Lambda)$ must be larger to provide evidence against the null.
- For $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$, the numerator of the likelihood ratio is the density function at point μ_0 and the denominator we plug in the mle of the parameter μ, (so here for example, $\bar{X}$). This follows from the definition, since the maximum of the likelihood function over the parameter space is the MLE.
- Knowing the null distribution of the test statistic makes possible the construction of a rejection region for any significance level α. Using the chi-square, RR is given by $\left|\bar{X} - \mu_0\right| \geq \frac{\sigma}{\sqrt{n}} z(\alpha/2)$ for χ^2_1, otherwise use chi square tables for more Df.
- Under smoothness conditions on the probability density or frequency functions involved, the null distribution of $-2log\Lambda$ tends to a chi-square distribution with degrees of freedom equal to $dim\,\Omega - dim\,\omega_0$ as the sample size tends to infinity.
- Generally - set up the distributions, compose the likelihood ratio, often take the log of both sides, try to determine how the likelihood changes for different observed values.
- $-2\log\Lambda \approx \sum_{i=1}^{m} \frac{\left[x_i - np_i(\hat{\theta})\right]^2}{np_i(\hat{\theta})} = X^2$: RHS is Pearson's test statistic for goodness of fit. Degrees of freedom (df): # of free parameters (think: # of unknown parameters)

Pearson Chi-Square Test

- Used on grouped data into bins
- $X^2 = \sum \frac{(Observed - Expected)^2}{Expected}$
- $X^2 := \sum_1^n \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^m \frac{\left[x_i - np_i(\hat{\theta})\right]^2}{np_i(\hat{\theta})} \sim \chi^2_{\mathrm{d}f:=n-s-1}$
- $x^2 = \text{ Pearson's cumulative test statistic, which asymptotically approaches a } \chi^2 \text{ distribution.}$
  $O_i = \text{ the number of observations of type } i.$
  $N = \text{ total number of observations}$
  $E_i = Np_i = \text{ the expected (theoretical) count of type } i, \text{ asserted by the null hypothesis that the fraction of type in the population is } p_i$
  $n = \text{ the number of cells in the table.}$
- Goodness-of-fit: whether eCDF differs from *any* theoretical CDF

1. Build test statistic for n bins - Oi is the observed count in each bin and Ei is the expected count in each bin
2. Null hypothesis is O ~ E
3. Compare the test statistic to critical values in the chi-square distribution
4. Degrees of freedom is number of cells less parameters being estimated.

Poisson Dispersion Test

- If one has a specific alternative hypothesis in mind, better power can usually be obtained by testing against that alternative rather than against a more general alternative.
- The two key assumptions underlying the Poisson distribution are that the rate is constant and that the counts in one interval of time or space are independent of the counts in disjoint intervals. These conditions are often not met.
- Given counts $x_1, \ldots, x_n$, we consider testing the null hypothesis that the counts are Poisson with the common parameter λ versus the alternative hypothesis that they are Poisson but have different rates, $\lambda 1, \ldots, \lambda_n$.
- $\Lambda = \frac{\prod_{i=1}^n \hat{\lambda}^{x_i} e^{-\hat{\lambda}}/x_i!}{\prod_{i=1}^n \tilde{\lambda}_i^{x_i} e^{-\tilde{\lambda}_i}/x_i!} = \prod_{i=1}^n \left(\frac{\bar{x}}{x_i}\right)^{x_i} e^{x_i - \bar{x}}$
- $-2\log\Lambda \approx \frac{1}{\bar{x}} \sum_{i=1}^n (x_i - \bar{x})^2$ using Taylor approximation.
- We use the above formula as the test statistic and find the relevant significance level / p-value from the poisson distribution.

Hanging Rootograms

- Hanging rootograms are a graphical display of the differences between observed and fitted values in histograms.
- Suppose we estimate $\bar{x} \to \mu$, $\hat{\sigma} \to \sigma$, the then probability that an observation falls in an interval between x_{j-1}, x_j is $\hat{p}_j = \Phi\left(\frac{x_j - \bar{x}}{\hat{\sigma}}\right) - \Phi\left(\frac{x_{j-1} - \bar{x}}{\hat{\sigma}}\right)$. For sample size n, the predicted count in the jth interval is $\hat{n}_j = n\hat{p}_j$, which we then compare to the observed counts
- The hanging histogram is then the difference between the fitted frequency of the bins and the observed. Expect larger fluctuations in the center than in the tails since variance across buckets is not constant - can use a variance-stabilizing transformation. Often use $f(x) = \sqrt{x}$ so get a hanging rootogram showing $\sqrt{n_j} - \sqrt{\hat{n}_j}$
- From delta-method, for $Y = f(X)$, $\mathrm{Var}(Y) \approx \sigma^2(\mu)[f'(\mu)]^2$. Variance is stabilized by making a transformation under which this is constant.
- Generally, deviations in the center have been down-weighted and those in the tails emphasized by the transformation.
- Hanging chi-gram: plots $\frac{n_j - \hat{n}_j}{\sqrt{\hat{n}_j}}$

Probability Plots

- Useful graphical tool for qualitatively assessing the fit of data to a theoretical distribution
- Plotting the ordered observations against expected values ($E\left(X_{(j)}\right) = \frac{j}{n+1}$) (find order statistic of data and plot against quantiles of a known distribution)
- Probability integral transform: $Y = F_X(X)$ to get uniform distribution. Then can plot against uniform quantiles.

Tests for Normality

- A goodness-of-fit test can be based on the coefficient of skewness: $b_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^3}{s^3}$, s is sample SD. Negative skew is left-sided: left tail is longer than right tail.
- The test rejects for large values of |b1|
- Coefficient of kurtosis: $b_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^4}{s^4}$. Normal has kurtosis of 3, so could have null hypothesis kurtosis = 3 and reject for large values of $|b_2 - 3|$.
- A goodness-of-fit test may also be based on the linearity of the probability plot, as measured by the correlation coefficient, r, of the x and y coordinates of the points of the probability plot. The test rejects for small values of r.
- Kolmogorov Smirnov Test - $D_n = \max_x |F_n(x) - F(x)|$ - largest vertical distance between the eCDF and the CDF.

Chapter 10 - Summarizing Data

CDF Methods

- empirical CDF: $F_n(x) = \frac{1}{n}(\#x_i \leq x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}\left(X_i\right)$ for a batch of numbers X, then just a count of numbers less than a specified value over the total batch size. Note indicators are Bernoullis: $I_{(-\infty,x]}\left(X_i\right) = \begin{cases} 1, & \text{with probability } F(x) \\ 0, & \text{with probability } 1 - F(x) \end{cases}$
  - $nF_n(x)$ is a binomial RV - n trials with F(x) probability of success $E\left[F_n(x)\right] = F(x)$
  - $\mathrm{Var}[F_n(x)] = \frac{1}{n}F(x)[1 - F(x)]$
  - Right continuous
- Survival Function: $S(t) = P(T > t) = 1 - F(t)$
  - Simply a reversal of the CDF for data consist of time until death or failure, chance of surviving past t.

Hazard Functions

- As the instantaneous death rate for individuals who have survived up to a given time. $h(t) = \frac{f(t)}{1-F(t)} = \frac{f(t)}{S(t)} = -\frac{d}{dt}log(S(t))$
- May be thought of as the instantaneous rate of mortality for an individual alive at time t. If T is the lifetime of a manufactured component, it may be natural to think of h(t) as the instantaneous or age-specific failure rate.

QQ Plots

- Shoes the ith order statistic $X_{(i)}$ against $F^{-1}(i/(n+1))$ for some distribution CDF F
- Pth quantile: $F(x) = p$ or $x_p = F^{-1}(p)$
- Plot the quantiles of one distribution against another
- Additive: if $y_p = x_p + h$, then for the quantiles $G(y) = F(y - h)$, for control group x with CDF F and treatment y with CDF G. For values of y G(y) = F(y) shifted to the right.
- Multiplicative: if $y_p = cx_p$, then for quantiles $G(y) = F(y/c)$
- If the values in the right tail would have to move down to the line that means they would have to become smaller and the data has a heavier tail than the distribution.

- To compare two batches of n numbers with order statistics X(1) ,..., X(n) and Y(1) ,..., Y(n) , a Q-Q plot is simply constructed by plotting the points (X(i) , Y(i)). The difference from PP plots is simply that we are plotting two data quantiles now instead of data against a theoretical distribution.

Histograms, Density Curves, Stem / Leaf plots

- Let w(x) be a weight function:
    - nonnegative
    - symmetric
    - centered at zero
    - integrating to 1
- We rescale w: $w_h(x) = \frac{1}{h} w\left(\frac{x}{h}\right)$. Small h causes kernel function to be more peaked about 0, large h more spread out.
- Kernel probability density estimate: $f_h(x) = \frac{1}{n}\sum_{i=1}^{n} w_h\left(x - X_i\right)$ - superposition of hills centered on the observations. If $w_h(x)$ is standard normal $w_h(x - Xi)$ is normal with mean Xi and SD h.
- The parameter h, bandwidth, controls the smoothness and is the bin width of the histogram. With histograms and density estimates, we lose information and cannot reconstruct the data.
- Stem and leaf plots - retain numerical information while showing shape.

Measures of Location

- Measure of the center of batch of numbers
- Arithmetic mean - sum over the count
- Robust measures - insensitive to outliers, such as the median.
- When the data are a sample from a continuous probability law, the sample median can be viewed as an estimate of the population median. The distribution of the number of observations greater than the median is binomial with n trials and probability 1/2 of success on each trial.
- Trimmed mean - The $100\alpha\%$ trimmed mean is easy to calculate: Order the data, discard the lowest $100\alpha\%$ and the highest $100\alpha\%$, and take the arithmetic mean of the remaining data: $\bar{x}_\alpha = \frac{x_{([n\alpha]+1)} + \cdots + x_{(n-[n\alpha])}}{n - 2[n\alpha]}$
- M Estimates - minimizers of $\sum_{i=1}^{n} \Psi\left(\frac{X_i - \mu}{\sigma}\right)$. where the weight function is a compromise between the weight functions for the mean and the median. Can be piecewise so say cutoff point of k, quadratic near zero and linaer beyond k reducing the influence of points beyond a certain bound.
- Estimating variability of location estimates by bootstrap:
    - Suppose we denote the location estimate as $\hat{\theta}$; it is important to keep in mind that $\hat{\theta}$ is a function of the random variables $X_1, X_2, \ldots, X_n$ and hence has a probability distribution, its sampling distribution, which is determined by n and F. We don't know F and $\hat{\theta}$ may be complicated.
    - We generate many samples from F (if we knew it) and calculate the value of $\hat{\theta}$, then could find measures on the samples like SD. Use empirical CDF instead as an approximation of F.
    - A sample of size n from Fn is thus a sample of size n drawn with replacement from the collection $x_1, x_2, \ldots, x_n$. We thus draw B samples of size n with replacement from the observed data, producing $\theta_1^*, \theta_2^*, \ldots, \theta_B^*$.
    - Then the SD is estimated as: $s_{\hat{\theta}} = \sqrt{\frac{1}{B}\sum_{i=1}^{B}\left(\theta_i^* - \bar{\theta}^*\right)^2}$

Measures of Dispersion

- Sample standard deviation: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ (use n-1 as divisor to make s^2 unbiased estimate of population variance)
    - If sample from standard normal, $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$
- Median absolute deviation from the median (MAD): the data are $x_1, \ldots, x_n$ with median $\tilde{x}$, the MAD is defined to be the median of the numbers $|x_i - \tilde{x}|$.
- These two measures of dispersion, the IQR and the MAD, can be converted into estimates of σ for a normal distribution by dividing them by 1.35 and .675, respectively.

Boxplot

- The range lines - A vertical line is drawn up from the upper quartile to the most extreme data point that is within a distance of 1.5 (IQR) of the upper quartile, ie $X_i > Q^{0.75} + 1.5 \times IQR$ and $X_i < Q^{0.25} - 1.5 \times IQR$. A similarly defined vertical line is drawn down from the lower quartile. Short horizontal lines are added to mark the ends of these vertical lines.

Chapter 11 - Comparing Two Samples

Independent Samples

- For independent samples - think individuals assigned to treatment vs control.
- Two sample shift model: $X_1, X_2, \ldots, X_{n_1}$ is a random sample from distribution F(X) and $Y_1, Y_2, \ldots, Y_{n_2}$ is a random sample from $G(y) = F(y - \theta)$ for an unknown θ. The null is that the distributions are equal, ie $\theta = 0$

Normal Methods

- The observations from the control group are modeled as independent random variables with a common distribution, F, and the observations from the treatment group are modeled as being independent of each other and of the controls and as having their own common distribution function, G.
- We will assume that a sample, $X_1, \ldots, X_n$, is drawn from a normal distribution that has mean μ_X and variance σ^2 , and that an independent sample, $Y_1, \ldots, Y_m$, is drawn from another normal distribution that has mean μ_Y and the same variance, σ^2
- For known population variance: estimate difference in means: $\bar{X} - \bar{Y} \sim N\left[\mu_X - \mu_Y, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right]$, with CI for known variance: $(\bar{X} - \bar{Y}) \pm z(\alpha/2)\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}$. Simply the statistic plus minus a multiple of its standard deviation
- Usually do not know variance, use pooled sample variance; $s_p^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{m+n-2}$ for $s_X^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$
- Test statistic: $t = \frac{(\bar{X}-\bar{Y}) - (\mu_X - \mu_Y)}{s_p\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim t_{m+n-2}$. Standard error of $\bar{X} - \bar{Y}$: $s_{\bar{X}-\bar{Y}} = s_p\sqrt{\frac{1}{n} + \frac{1}{m}}$. Confidence interval: $(\bar{X} - \bar{Y}) \pm t_{m+n-2}(\alpha/2)s_{\bar{X}-\bar{Y}}$

- Hypothesis testing: $H_0 : \mu_X = \mu_Y$, $H_1 : \mu_X \neq \mu_Y$. Uses test stat: $t = \frac{\bar{X}-\bar{Y}}{s_{\bar{X}-\bar{Y}}}$ since essentially testing if the difference has zero mean. If we wanted to test against a non-zero difference, might look something like $\mu_X - (\mu_Y + 5) = 0$ This test can be derived from the GLRT - rejects for large values.
- Without assumption of equal variances, estimate of $\mathrm{Var}(\bar{X} - \bar{Y}) = \frac{s_X^2}{n} + \frac{s_Y^2}{m}$ and we use the t-distribution with $\mathrm{df} = \frac{\left[(s_X^2/n)+(s_Y^2/m)\right]^2}{\frac{\left(s_X^2/n\right)^2}{n-1}+\frac{\left(s_Y^2/m\right)^2}{m-1}}$
- Procedure:
    - Determine distribution of $\bar{X} - \bar{Y}$ and build normalized test statistic Z.
    - If σ is unknown, calculate s_p and build test statistic t ($t = \frac{\bar{X}-\bar{Y}}{s\bar{x}-\bar{Y}}$ for $H_0 = 0$). If variances are unequal, $\mathrm{Var}(\bar{X} - \bar{Y}) = \frac{s_X^2}{n} + \frac{s_Y^2}{m}$ and modify t distribution degrees of freedom
    - Hypothesis test against null or $(\bar{X} - \bar{Y}) \pm t_{m+n-2}(\alpha/2)s_{\bar{X}-\bar{Y}}$
    - Power:
        - For alternative hypothesis $H_1 : \mu_X - \mu_Y = \Delta$, calculate power using the normal distribution (since special noncentral t tables would be need otherwise).
        - Then RHS power is given by $1 - \Phi\left[z(\alpha/2) - \frac{\Delta}{\sigma}\sqrt{\frac{n}{2}}\right]$ ie. the (z stat used for the RR) - (mean difference)/(pooled SE).
        - Total two tailed power: $1 - \Phi\left[z(\alpha/2) - \frac{\Delta}{\sigma}\sqrt{\frac{n}{2}}\right] + \Phi\left[-z(\alpha/2) - \frac{\Delta}{\sigma}\sqrt{\frac{n}{2}}\right]$
- It is sometimes advocated that skewed data be transformed to a more symmetric shape before normal theory is applied. Transformations such as taking the log or the square root can be effective in symmetrizing skewed distributions because they spread out small values and compress large ones.
- The ratio of the standard deviation of a distribution to the mean is called the coefficient of variation (CV); it expresses the standard deviation as a fraction of the mean.
- The power of the two-sample t test depends on four factors:
    1. The real difference $\triangle = |\mu_X - \mu_Y|$. The larger this difference, the greater the power.
    2. The significance level α at which the test is done. Large alpha larger power
    3. The smaller the population standard deviation, the larger the power.
    4. The sample sizes n and m. The larger the sample sizes, the greater the power.

Nonparametric Method - Mann-Whitney

- Suppose that we have m + n experimental units to assign to a treatment group and a control group. The assignment is made at random: n units are randomly chosen and assigned to the control, and the remaining m units are assigned to the treatment.
- First, we group all m + n observations together and rank them in order of increasing size. We next calculate the sum of the ranks of those observations that came from the control group. If this sum is too small or too large, we will reject the null hypothesis. We consult a rank table to determine the level of significance for the rank sum obtained for the group with the smaller rank sum. We have not made any assumption that the observations from the control and treatment groups are samples from a probability distribution.

- When it is more appropriate to model the control values, $X_1, \ldots, X_n$, as a sample from some probability distribution F and the experimental values, $Y_1, \ldots, Y_m$, as a sample from some distribution G, the Mann-Whitney test is a test of the null hypothesis $H_0 : F = G$. The reasoning is exactly the same: Under H_0, any assignment of ranks to the pooled m + n observations is equally likely.
- If the groups are roughly the same, then the ranks will have a good amount of alternation and be mixed from each group. The more separation in the ranks, the more different the two groups.
- $U = nm + \frac{n(n+1)}{2} - W$, for W = rank sum for sample X. U is obtained by ordering all n+m observations and counting the number of observations in sample I that precede each observation in sample II - U is the sum of these counts
- Let T_Y denote the sum of the ranks of $Y_1, Y_2, \ldots, Y_m$. $E(T_Y)$ and $Var(T_Y)$ under the null hypothesis F = G: $E\left(T_Y\right) = \frac{m(m+n+1)}{2}$, $\operatorname{Var}(T_Y) = \frac{mn(m+n+1)}{12}$
  - Under $H_0 : F = G$, $E\left(U_Y\right) = \frac{mn}{2}$, $\operatorname{Var}(U_Y) = \frac{mn(m+n+1)}{12}$, and $\frac{U_Y - E(U_Y)}{\sqrt{\operatorname{Var}(U_Y)}} \sim N(0,1)$ for m, n over 10.
- Since the actual numerical values are replaced by their ranks, the test is insensitive to outliers, whereas the t test is sensitive. It has been shown that even when the assumption of normality holds, the Mann-Whitney test is nearly as powerful as the t test and it is thus generally preferable, especially for small sample sizes.
- Bootstrap: As before, suppose that $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ are two independent samples from distributions F and G, respectively, and that $pi = P(X < Y)$ is estimated by $\hat{\pi}$. How can the standard error of $\hat{\pi}$ be estimated and how can an approximate confidence interval for π be constructed?
  - An approximation can be obtained by using the empirical distributions F_n and G_n in their places. This means that a bootstrap value of $\hat{\pi}$ is generated by randomly selecting n values from $X_1, X_2, \ldots, X_n$ with replacement, m values from $Y_1, Y_2, \ldots, Y_m$ with replacement and calculating the resulting value of $\hat{\pi}$
- Procedure
  - No assumed distribution:
    - Group all m + n observations together and rank them in order of increasing size
    - Calculate the sum of the ranks of those observations. If R_1 sum of one sample, then the other sample has rank sum $R_2 = n_1(m + n + 1) - R_1$ where n_1 is the smaller sample size. Then use the smaller statistic
    - Under the null, every assignment of m+n ranks to observations is equally likely, hence each of the $\binom{m+n}{m}$ assignments to the control group is equally likely.
    - For large m,n rank sums are approx normal: $\frac{R_1 - \mathbb{E}R_1}{\sigma_1} \sim \Phi$ for $\mathbb{E}\mathcal{R}_1 = \frac{m(m+n+1)}{2}$ and $\mathbb{V}\mathcal{R}_1 = \frac{mn(m+n+1)}{12}$
  - Test statistic: $U_1 = R_1 - \frac{n_1(n_1+1)}{2}$ for n_1 sample size of of sample 1 and R_1 sum of ranks in sample 1
  - When we assume a distribution for control variables $X \; simF$ and experimental variables $Y \sim G$, M-W is a test of null $H_0 : F = G$
  - We can for larger samples normalize the test stat sum of ranks T $\frac{T - E(T)}{\sigma_T}$ to use normal distribution

- CI's in a Shift Model $G(x) = F(x - \Delta)$

Bayesian Approach

- The X_i are i.i.d. normal with mean μ_X and precision ξ; and the Y_j are i.i.d. normal with mean μ_Y, precision ξ, and independent of the X_i
- We use improprer priors to get an approx result
- $\frac{\Delta - (\bar{x} - \bar{y})}{s_p \sqrt{n^{-1} + m^{-1}}} \sim t_{n+m-2}$ but here the mean difference and s_p are fixed and $\triangle$ is random. The posterior can then be found using the t distribution: $= P\left(T \geq \frac{\bar{y} - \bar{x}}{s_p \sqrt{n^{-1} + m^{-1}}}\right)$

Comparing Paired Samples

- May match subjects with related characteristics, then assign one to the control and one to the experimental group. The variance of a paired sample will be smaller due to the correlation term.
- A pair is (X_i, Y_i) and these are not assumed to be independent (think left and right hand strength), but for $i \neq j$, (X_i, X_j) can be independent, think left hand strength across a population.
  - sampling two different people: $X_i - Y_j \sim N\left(\mu_x - \mu_u, \sigma_x^2 + \sigma_y^2\right)$
  - sampling one person: $X_i - Y_i \sim N\left(\mu_x - \mu_u, \sigma_x^2 + \sigma_y^2 - 2\rho\sigma_x\sigma_y\right)$
- Pairs are (X_i, Y_i), $i = 1, \ldots, n$, different means and variances. Different pairs are independently distributed and $\mathrm{Cov}(X_i, Y_i) = \sigma_{XY}$
- For $D_i = X_i - Y_i$, independent with $E(D_i) = \mu_X - \mu_Y$,
  $\mathrm{Var}(D_i) = \sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY} = \sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y$

Normal Methods

- $E(D_i) = \mu_X - \mu_Y = \mu_D$, $\mathrm{Var}(D_i) = \sigma_D^2$
- $D_i \sim N\left(\mu_x - \mu_y, \sigma_x^2 + \sigma_y^2 - 2\rho\sigma_x\sigma_y\right) = N\left(\mu_D, \sigma_D^2\right)$
- Test stat: $t = \frac{\bar{D} - \mu_D}{s_{\bar{D}}}$, CI: $\bar{D} \pm t_{n-1}(\alpha/2)s_{\bar{D}}$, Two sided RR: $|\bar{D}| > t_{n-1}(\alpha/2)s_{\bar{D}}$
- Procedure
  - Calculate differences in paired samples - say before and after a treatment for each patient: $\bar{D}$
  - Build test statistic $t = \frac{\bar{D} - \mu_D}{s_{\bar{D}}}$, use t_{n-1} to find critical value

Nonparametric Method - Sign Test

- n pairs of observations (X_i, Y_i), null is that the distributions of X and Y are the same. The probability that the mean difference $D_i > 0 = 0.5$ If M is the total number of positive differences, under the null $M \sim Bin(n, 1/2)$

1. Let $p = P(X > Y)$
2. Null hypothesis: $H_0 : p = 1/2$ Alternative hypothesis: $H_a : p > 1/2$ or $(p < 1/2$ or $p \neq 1/2)$
3. Test statistic: M = # of positive $D_i = X_i - Y_i$
4. Rejection region for $H_a : p > 1/2$, then reject for largest values of M
5. Assumptions: pairs are randomly and independently selected

- When we encounter ties due to equal observations in pairs, can delete and reduce n. If n is large, can use normal approximation $Z = \frac{M - np}{\sqrt{npq}}$

Nonparametric Method - Wilcoxon Signed Rank Test

- Under the null of equal distributions, if we were to order the differences according to their absolute values and rank them from smallest to largest, the expected rank sums for the negative and positive differences would be equal.
- Non parametric version of the paired sample t test and does not depend on normality. Also insensitive to outliers since uses ranks. Preferable for small samples since nearly has the power of the t-test with normal samples too.
- Procedure
  - Calculate differences D_i, removing differences equal to 0. Under the null D is symmetric about 0
  - Rank by absolute values of the diffs, averaging the ranks for tied differences
  - Restore the signs of the differences to the ranks
  - Calculate W_+, the sum of ranks that have positive signs. Use the smaller sum as our test statistic - the smaller the value of the statistic, the greater the weight of evidence favoring rejection. Use W- to detect shifts of Y to the left of X; W+ to detect shifts of Y to the right of X
  - For two tailed test $T = \min(T^+, T^-)$. Reject null H_0 if $T \leq T_0$ for CV of two tailed test.
  - For larger samples, a normal approximation of the null distribution can be used, using the expectation and variance below.
- If one condition produces larger values than the other, W_+ will take on extreme values. We test the null hypothesis that the distribution of D_i is symmetric about zero. Equivalent then to Bern(0.5).
- For normal approximation with sample size greater than 20:
  $E(W_+) = \frac{n(n+1)}{4}$, $Var(W_+) = \frac{n(n+1)(2n+1)}{24}$. For large samples $z = \frac{T^+ - E(T^+)}{\sqrt{V(T^+)}} = \frac{T^+ - [n(n+1)/4]}{\sqrt{n(n+1)(2n+1)/24}}$

Chapter 12 - ANOVA

One Way Layout

- An experimental design in which independent measurements are made under each of several treatments. Analogous to the two independent samples tests in chapter 11. Extending Chapter 11 to many samples

Normal Theory / F-Test

- We first discuss the analysis of variance and the F test in the case of **I groups, each containing J samples**. The I groups will be referred to generically as treatments, or levels
- Let $Y_{ij} =$ the jth observation of the ith treatment
- $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$
  - observations are corrupted by random independent errors ϵ_{ij}, normally distributed with mean zero and constant variance σ^2
  - F test is approximately valid for large non-normal samples
  - μ is the overall mean level
  - α_i is the differential effect of the ith treatment normalized st $\sum_{i=1}^{l} \alpha_i = 0$.
- Expected response to ith treatment: $E(Y_{ij}) = \mu + \alpha_i$

- The total sum of squares equals the sum of squares within groups plus the sum of square between groups:
  $SS_{TOT} = SS_W + SS_B \rightarrow \sum_{i=1}^{I}\sum_{j=1}^{J}\left(Y_{ij} - \bar{Y}_{..}\right)^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(Y_{ij} - \bar{Y}_{i.}\right)^2 + J\sum_{i=1}^{I}\left(\bar{Y}_{i.} - \bar{Y}_{..}\right)^2$
- Lemma A: For $X_i \perp$ RVs with means μ_i and shared variance σ^2 then
  $E\left(X_i - \bar{X}\right)^2 = (\mu_i - \bar{\mu})^2 + \frac{n-1}{n}\sigma^2$
- Then Theorem A: $E\left(SS_W\right) = \sum_{i=1}^{I}\sum_{j=1}^{J} E\left(Y_{ij} - \bar{Y}_{i.}\right)^2 = I(J-1)\sigma^2$.
  - Can use for unbiased estimate of σ^2 : $s_p^2 = \frac{SS_w}{I(J-1)}$, $SS_W = \sum_{i=1}^{I}(J-1)s_i^2$
  - If all α_i are zero, then expected SSw and SSb should be about equal, if some alphas are not zero then SSb increases.
- Theorem B: For independent, N(0, σ^2) errors, $SS_W/\sigma^2 \sim \chi^2_{I(J-1)}$. If $\alpha_i = 0\,\forall i$, $SS_B/\sigma^2 \sim \chi^2_{I-1}$ and $SS_W \perp SS_B$
- Procedure
  - We use test statistic $F = \frac{SS_B/(I-1)}{SS_W/[I(J-1)]}$ to test $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_l = 0$. The denominator of the F statistic has expected value equal to σ^2 , and the expectation of the numerator is $J(I-1)^{-1}\sum_{i=1}^{I}\alpha_i^2 + \sigma^2 2$. Thus, if the null hypothesis is true, the F statistic should be close to 1, whereas if it is false, the statistic should be larger. Reject hypothesis for large values of F.
  - Under normally distributed errors, null distribution of $F \sim F_{(I-1),(I(J-1))}$
- In ANOVA table,
  - SS due to Labs = SSb
  - SS due to error = SSw
  - Mean square is the sum of squares divided by the Df (I-1 or I(J-1))
- For unequal number of observations under various treatments:
  - $E\left(SS_w\right) = \sigma^2\sum_{i=1}^{I}\left(J_i - 1\right)$, $E\left(SS_B\right) = (I-1)\sigma^2 + \sum_{i=1}^{I} J_i\alpha_i^2$

Kruskal-Wallis Test

- Generalization of Mann-Whitney test for independent observations with no assumed distributions
- We assume that independent random samples have been drawn from k populations that differ only in location. The samples sizes may be unequal, and n_i is the sample size drawn from the ith population. Combine all samples into n observations and rank them from smallest to largest. Tied ranks are averaged.
- R_i = sum of ranks for observations from population i, $\bar{R}_i = R_i/n_i$ is the average of the ranks. If the null hypothesis is true and the populations do not differ in location, we would expect the $\bar{R}_i$ values to be approximately equal and $V = \sum_{i=1}^{k} n_i\left(\bar{R}_i - \bar{R}\right)^2 = \sum_{i=1}^{k} n_i\left(\bar{R}_i - \frac{n+1}{2}\right)^2$ should be small.
- K-W use test statistic $H = \frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i} - 3(n+1)$. Then $P[H > h(\alpha)] = \alpha$. For large enough n_i (say 5), $H \sim \chi^2_{k-1}$. Then we reject H_0 if $H > \chi^2_\alpha$ with $(k-1)$ df.
- Using Rice notation:
  - $R_{ij} =$ = the rank of Y_{ij} in the combined sample
  - $\bar{R}_{i.} = \frac{1}{J_i}\sum_{j=1}^{J_i} R_{ij}$,
  - $\bar{R}_{..} = \frac{N+1}{2}$

- $SS_B = \sum_{i=1}^{I} J_i \left(\bar{R}_{i.} - \bar{R}_{..}\right)^2$
- $K = \frac{12}{N(N+1)} \left(\sum_{i=1}^{I} J_i \bar{R}_{i.}^2\right) - 3(N+1)$

- Procedure
  - Pool observations and rank, letting $R_{ij} =$ the rank of Y_{ij} in the combined sample
  - SS_B is a measure of dispersion of the $\bar{R}_{i.}$ - used to test the null that the distributions generating the observations under the treatments are identical.
  - For hand computation: test statistic $K = \frac{12}{N(N+1)} \left(\sum_{i=1}^{I} J_i \bar{R}_{i.}^2\right) - 3(N+1) \sim \chi^2_{I-1}$

Problem of Multiple Comparisons

- Real interest may be focused on comparing pairs or groups of treatments and estimating the treatment means and their differences - the F-test does not tell us how our treatment effects differ.
- Tukey's Method - to construct confidence intervals for the differences of all pairs of means in such a way that the intervals simultaneously have a set coverage probability, then rely on the duality of tests and CIs
  - $\max_{i_1,i_2} \frac{\left|\left(\bar{Y}_{i_1\cdot} - \mu_{i_1}\right) - \left(\bar{Y}_{i_2\cdot} - \mu_{i_2}\right)\right|}{s_p/\sqrt{J}}$ random variable where max is taken over all pairs. Called the studentized range distribution with parameters I (number of samples being compared) and I(J-1) (df of s_p).
  - CI constructed as $\left(\bar{Y}_{i_1\cdot} - \bar{Y}_{i_2\cdot}\right) \pm q_{I,I(J-1)}(\alpha)\frac{s_p}{\sqrt{J}}$, reject if $\left|\bar{Y}_{i_1\cdot} - \bar{Y}_{i_2\cdot}\right| > q_{I,I(J-1)}(\alpha)\frac{s_p}{\sqrt{J}}$
  - Essentially calculate $q_{I,I(J-1)}(\alpha)\frac{s_p}{\sqrt{J}}$ and see if there are differences in means of different labs that exceed this value. q is a percentile of the studentized range and s_p is the square root of the mean square for error in the anova table - $\sqrt{SS_W/I(J-1)}$
- Bonferroni Method: Desired error rate α can be obtained over k null hypotheses by testing each null hypothesis at level α/k. Advantage of not needing the sample sizes for each treatment.
  - k is the number of pairwise comparisons $\binom{I}{2}$

Two Way Layout

- A two-way layout is an experimental design involving two factors, each at two or more levels. The levels of one factor might be various drugs, for example, and the levels of the other factor might be genders. If there are **I levels of one factor and J of the other**, there are I × J combinations. Assume **K independent observations** are taken for each of these combinations
- Think of 3 electric ranges cooking over 3 menus - 9 total means to compare

Additive Parametrization

- $\hat{Y}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j$
- Calculate total average. Calculate the means for each I group across J. Differential effects $\hat{\alpha_i}$ are the differences between the group means and overall mean. Calculate the means for each group J across I, and repeat the calculation of differential effects $\hat{\beta}_j$.
- The differences of the observed values and the fitted values, $Y_{ij} - \hat{Y}_{ij}$ are the residuals from the additive model. $\sum_{i=1}^{3} \hat{\delta}_{ij} = \sum_{j=1}^{3} \hat{\delta}_{ij} = 0$ for residuals $\delta_{ij} = Y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j$. Then the model $Y_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\delta}_{ij}$ fits the data points exactly.

Normal Theory for Two-Way

- Balanced design - equal number of observations per cell, and we assume K > 1 per cell
- $Y_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \varepsilon_{ijk}$ for Y_{ijk}, the kth observation in cell ij. ϵ_{ijk} are random errors iid $N(0, \sigma^2)$ (common variance).
- Therefore $E(Y_{ijk}) = \mu + \alpha_i + \beta_j + \delta_{ij}$ and the parameters are constrained by $\sum_{i=1}^{I} \alpha_i = 0$, $\sum_{j=1}^{J} \beta_j = 0$, $\sum_{i=1}^{I} \delta_{ij} = \sum_{j=1}^{J} \delta_{ij} = 0$
- Use MLE on the unknown parameters. Comparing sum of squares we get $SS_{TOT} = SS_A + SS_B + SS_{AB} + SS_E$
  - Under error assumptions above:
  - $E(SS_A) = (I-1)\sigma^2 + JK\sum_{i=1}^{I} \alpha_i^2$
  - $E(SS_B) = (J-1)\sigma^2 + IK\sum_{j=1}^{J} \beta_j^2$
  - $E(SS_{AB}) = (I-1)(J-1)\sigma^2 + K\sum_{i=1}^{I}\sum_{j=1}^{J} \delta_{ij}^2$
  - $E(SS_E) = IJ(K-1)\sigma^2$
- $SS_E/\sigma^2 \sim \chi^2_{IJ(K-1)}$
- Under null $H_A : \alpha_i = 0, i = 1, \ldots, I$, $SS_A/\sigma^2 \sim \chi^2_{I-1}$
- Under null $H_B : \beta_j = 0, \quad j = 1, \ldots, J$, $SS_B/\sigma^2 \sim \chi^2_{J-1}$
- Under null $H_{AB} : \delta_{ij} = 0, i = 1, \ldots, I, j = 1, \ldots, J$, $SS_{AB}/\sigma^2 \sim \chi^2_{(I-1)(J-1)}$
- F tests of these null hypotheses are conducted by comparing the appropriate SS to the sum of squares for error.
-

Analysis of Variance Table

Source	*df*	*SS*	*MS*	*F*
Iron form	1	2.074	2.074	5.99
Dosage	2	15.588	7.794	22.53
Interaction	2	.810	.405	1.17
Error	102	35.296	.346	
Total	107	53.768		

- In the following analysis of variance table, SS A is the sum of squares due to the form of iron, SS B is the sum of squares due to dosage, and SS AB is the sum of squares due to interaction. The F statistics were found by dividing the appropriate mean square by the mean square for error.
- Eg. to test effect of form of iron, $H_A : \alpha_1 = \alpha_2 = 0, F = \frac{SS_{IRON}/1}{SS_E/102} = 5.99$

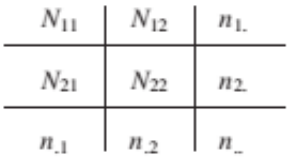
Chapter 13 - Analysis of Categorical Data

- Two way tables - suppose rows are hair colors and columns eye colors, each cell is a count of people who fall in that cross classification - can we find a relationship between hair color and eye color?

Fisher's Exact Test

- According to the null hypothesis, the margins of the table are fixed - ie the overall counts for each category. The randomization determines the counts in the interior of the table (capital letters) subject to

the margin constraints, leaving us with one degree of freedom.

- The row and columns totals are fixed by the experimental design and are not treated as random. Only the entries are are treated as random. Since the rows and columns are fixed, the entries all rely on a single random variable and have 1 degree of freedom.
- 

N_{11}	N_{12}	$n_{1.}$
N_{21}	N_{22}	$n_{2.}$
$n_{.1}$	$n_{.2}$	$n_{..}$

- Under the null N_{11} is distributed as the number of successes in 24 draws without replacement from a population of 35 successes and 13 failures - hypergeometric. The probability
  $N_{11} = n_{11} = p\left(n_{11}\right) = \frac{\binom{n_{1.}}{n_{11}}\binom{n_{2.}}{n_{21}}}{\binom{n_{..}}{n.1}}$
- We use N_{11} as the test statistic to test the null.

Chi-Square Test of Homogeneity

- Suppose that we have independent observations from **J multinomial distributions**, **each of which has I cells**, and that we want to test whether the cell probabilities of the multinomials are equal—that is, to test the homogeneity of the multinomial distributions.
- We fix the J column totals and treat each column as its own multinomial distribution. The question is whether those column distributions are the same.
- The set up here is we are testing I cells (the words in Austen's books) against a number of samples (the different books). J is simply a count of the books, so we do not treat the tests as having two moving parameters. We look at a given i (word) across the distributions (the books) to see if they have the same frequency of appearing.
- Example: how close is an admirer to matching Jane Austen's style using word counts in their works
- The six word counts for Sense and Sensibility will be modeled as a realization of a multinomial random variable with unknown cell probabilities and total count 375; the counts for the other works will be similarly modeled as independent multinomial random variables.
- Thus, we must consider comparing J multinomial distributions each having I categories. If the probability of the ith category of the jth multinomial is denoted π_{ij} , the null hypothesis to be tested is: $H_0 : \pi_{i1} = \pi_{i2} = \cdots = \pi_{iJ}, \quad i = 1, \ldots, I$. This is essentially a goodness of fit test
- Under H0, each of the J multinomials has the same probability for the ith category, say π_i . The following theorem shows that the mle of π_i is simply $ni. /n. ..$ Here, $n_{i.}$ is the total count in the ith category, $n_{..}$ is the grand total count, $n_{.j}$ is the total count for the jth multinomial.
- $E_{ij} = \frac{n_{.j}n_{i.}}{n_{..}}$ for the jth multinomial the expected count in the ith category is the estimated probability of that cell times the total number of observation for the jth multinomial. Can use Pearson Chi-Square: $X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(n_{ij}-n_{i.}n_{.j}/n_{..}\right)^2}{n_{i.}n_{.j}/n_{..}}$ with $df = (I-1)(J-1)$
- Calculate the expected count for each cell using the chi square distribution. Plug into Pearson test for all cells

Chi-Square Test of Independence

- Education vs marital status - is there a relationship? College / No College vs Married Once / Married more than once

- We fix the overall n - we have a fixed sample size. We drop balls from the sky and they land in an I by J grid. We have a single multinomial distribution with IJ levels and probability π_{ij} on the (i,j) bin.
- Here we are looking to see if our rows and columns are independent. Are marriage and education independent of each other, so we set up the null that the probability in a given cell is the overall i probability times the overall j probability, ie the number of observations in the i group (over all j's) times the number of observations of in the j group (across all i's). Calculated over all i and j. We treat the whole table as coming from a single multinomial distribution.
- We will discuss statistical analysis of a sample of size n cross-classified in a table with I rows and J columns. Such a configuration is called a **contingency table**. The joint distribution of the counts n_{ij} , where i = 1,..., I and j = 1,..., J , is multinomial with cell probabilities denoted as $\pi_i j$. Let $\pi_{i.} = \sum_{j=1}^{J} \pi_{ij},\ \pi_{.j} = \sum_{i=1}^{I} \pi_{ij}$, denote the marginal probabilities that an observation will fall in the ith row and jth column, respectively.
- If rows and columns are independent of each other, then $\pi_{ij} = \pi_{i.}\pi_{.j}$. Therefore the null hypoth is $H_{)} = \pi_{ij} = \pi_{i.}\pi_{.j}$ for all i,j versus the alternative that the π_{ij} are free. Under null, mle is $\hat{\pi}_{ij} = \hat{\pi}_{i.}\hat{\pi}_{.j} = \frac{n_{i.}}{n} \times \frac{n.\,j}{n}$. Under the alternative, mle is $\tilde{\pi}_{ij} = \frac{n_{ij}}{n}$.
- Turning to Pearson test, $E_{ij} = n\hat{\pi}_{ij} = \frac{n_{i.}n_{.j}}{n}$, so $X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - n_{i.}n_{.j}/n)^2}{n_{i.}n_{.j}/n}$, $df = (I-1)(J-1)$
- The chi-square statistic used here to test independence is identical in form and degrees of freedom to that used in the preceding section to test homogeneity; however, the hypotheses are different and the sampling schemes are different. The test of homogeneity was derived under the assumption that the column (or row) margins were fixed, and the test of independence was derived under the assumption that only the grand total was fixed. Independence can be thought of as homogeneity of conditional distributions; for example, if education level and marital status are independent, then the conditional probabilities of marital status given educational level are homogeneous
- The more items we treat as unknown in the model, the degrees of freedom increase. This will lower the power we will achieve and increase our p-values - our estimation will be more uncertain because we allow more to vary.

Matched-Pairs Designs

- The assumption behind the chi-square test of homogeneity is that independent multinomial samples are compared, and sibling samples are not independent, because siblings are paired. With positive covariance $\sigma_{XY} > 0$ between pairs in the sample, we get more power.
- The null hypothesis is that probabilities of outcomes are the same for the X component and Y component of the pair in each sample, ie $\pi_{1.} = \pi_{.1}$ and $\pi_{2.} = \pi_{.2}$ The null can be written as $H_0 : \pi_{12} = \pi_{21}$. The off diagonal probabilities are equal and under the alternative they are not.
π_{11}	π_{12}	$\pi_{1.}$
π_{21}	π_{22}	$\pi_{2.}$
$\pi_{.1}$	$\pi_{.2}$	1
- McNemer's Test: Under null MLEs of cell probabilities are $\hat{\pi}_{11} = \frac{n_{11}}{n},\ \hat{\pi}_{22} = \frac{n_{22}}{n},\ \hat{\pi}_{12} = \hat{\pi}_{21} = \frac{n_{12} + n_{21}}{2n}$
- The $n_{11},\ n_{22}$ contributions to chi square test are 0 - we do not care about the diagonals
- $X^2 = \frac{\left[n_{12} - (n_{12} + n_{21})/2\right]^2}{(n_{12} + n_{21}/2} + \frac{\left[n_{21} - (n_{12} + n_{21})/2\right]^2}{(n_{12} + n_{21})/2} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$ with 1 degree of freedom

Odds Ratios

- $\text{odds}(A) = \frac{P(A)}{1-P(A)}$ implying $P(A) = \frac{\text{odds}(A)}{1+\text{odds}(A)}$
- Now suppose that X denotes the event that an individual is exposed to a potentially harmful agent and that D denotes the event that the individual becomes diseased. We denote the complementary events as $\bar{X}$ and $\bar{D}$.
  - $\text{odds}(D|X) = \frac{P(D|X)}{1-P(D|X)}$ and $\text{odds}(D|\bar{X}) = \frac{P(D|\bar{X})}{1-P(D|\bar{X})}$
  - The odds ratio is then $\Delta = \frac{\text{odds}(D|X)}{\text{odds}(D|\bar{X})}$ and measures the influence of exposure on subsequent disease
- Odds ratio can be the product of diagonal probabilities in the table divided by the product of the off diagonal probabilities: $\Delta = \frac{\pi_{11}\pi_{00}}{\pi_{01}\pi_{10}}$
-

	$\overline{D}$	D	
$\overline{X}$	π_{00}	π_{01}	$\pi_{0.}$
X	π_{10}	π_{11}	$\pi_{1.}$
	$\pi_{.0}$	$\pi_{.1}$	1

- Prospective study: a fixed number of exposed and nonexposed individuals are sampled, and the incidences of disease in those two groups are compared. We can calculate the odds ratio but not the individual probabilities π_{ij} since the marginal counts have been fixed by the sample design
- Retrospective study: a fixed number of diseased and undiseased individuals are sampled and the incidences of exposure in the two groups are compared. The joint / conditional probabilities cannot be calculated, but we can say $\text{odds}(X|D) = \frac{\pi_{11}}{\pi_{01}}$, $\text{odds}(X|\bar{D}) = \frac{\pi_{10}}{\pi_{00}}$ leading to estimated odds ratio $\hat{\Delta} = \frac{n_{00}n_{11}}{n_{01}n_{10}}$ - essentially by what factor does exposure increase odds of disease
- Analytical derivation of standard error is difficult, often use bootstrap using a binomial model for N_{11} with n = $n_{00} + n_{01}$ and $p = \pi_{11}$.

Chapter 14 - Linear Least Squares

- Minimizing $S(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$ with betas chosen to minimize the sum of squared vertical deviations. Procedure is not symmetric is x and y since using vertical distance.
- To find betas
  - take partial derivatives of S wrt beta $\frac{\partial S}{\partial \beta_1} = -2\sum_1^n x_i(y_i - \beta_0 - \beta_1 x_i) = 0$, $\frac{\partial S}{\partial \beta_0} = -2\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0$
  - $\sum_{i=1}^n y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_i$ and $\sum_{i=1}^n x_i y_i = \hat{\beta}_0 \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2$
  - Set derivatives to zero, collect terms, divide by n
- If the function to be fit is not linear in the unknown parameters, a system of nonlinear equations must be solved to find the coefficients. Typically, the solution cannot be found in closed form, so an iterative procedure must be used.

Simple Linear Regression

- The least squares estimates are unbiased: $E(\hat{\beta}_j) = \beta_j$. Only depends on the errors being additive with 0 mean. Does not depend on errors having same variance and independent.
- From Theorem B, we see that the variances of the slope and intercept depend on the xi and on the error variance. The xi are known; therefore, to estimate the variance of the slope and intercept, we need to estimate only σ^2.
- Estimate variance through RSS: $\sum_1^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$, and $s^2 = \frac{RSS}{n-2}$ is an unbiased estimate of sigma squared
- For large n and independent errors, estimated betas are approximately normally distributed $\frac{\hat{\beta}_i - \beta_i}{s_{\hat{\beta}_i}} \sim t_{n-2}$, which allows us to use the t distribution to be used for CIs
- Residuals $\hat{e}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$
- The model assumes homoskedastic errors - otherwise standard errors and CIs based on s^2 will underestimate. Can transform via log or squareroot to stabilize the variance.
- For $s_{xx} = \frac{1}{n}\sum_1^n (x_i - \bar{x})^2$, $s_{xy} = \frac{1}{n}\sum_1^n (y_i - \bar{y})^2$, $s_{xy} = \frac{1}{n}\sum_1^n (x_i - \bar{x})(y_i - \bar{y})$
  - the correlation coefficient between x and y is $r = \frac{s_{xy}}{\sqrt{s_{xx} s_{yy}}}$,
  - $\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}} \rightarrow r = \hat{\beta}_1 \sqrt{\frac{s_{xx}}{s_{yy}}}$.
  - Therefore correlation is zero iff slope is zero
- Standard equations for coefficients: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ and $\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$
- $\frac{\hat{y} - \bar{y}}{\sqrt{s_{yy}}} = r \frac{\hat{x} - \bar{x}}{\sqrt{s_{xx}}}$ - for r greater than 0 and x is 1 SD avove its mean, then the predicted value of y is r SDs bigger than its average. The predicted value deviates from its average by fewer SDs r < 1 than the predictor.

Statistical Properties of Least Squares Estimates

- Under the assumption that the errors have mean zero, the least squares estimates are unbiased
- Under the assumption that the errors have mean zero and are uncorrelated with constant variance σ^2, the covariance matrix of the least squares estimate $\hat{\beta} = \Sigma_{\hat{\beta}\hat{\beta}} = \sigma^2 (X^T X)^{-1}$
- Under the assumption that the errors are uncorrelated with constant variance, an unbiased estimate of σ^2 is $s^2 = \frac{||Y - \hat{Y}||^2}{n-p}$
- CI for beta - $\hat{\beta}_i \pm t_{n-p}(\alpha/2) s_{\hat{\beta}_i}$

Multiple Regression

- Squared multiple correlation coefficient / coefficient of determination $R^2 = \frac{s_y^2 - s_{\hat{e}}^2}{s_y^2}$

CIs, Inference, Bootstrap

- Develop model in which X and Y are random, whereas before the X's were fixed with randomness only from errors.
- Design matrix = random matrix Ξ and a particular realization of it is X. Each ξ_i is a row in Ξ and rows in X are xi. Then the model is $E(Y|\xi = x) = x\beta$, $Var(Y|\xi = x) = \sigma^2$
- The previous model is a conditional model of this generalized model.
- The beta estimates are still unbiased but their variances are $Var(\hat{\beta}_i) = \sigma^2 E(\Xi^T \Xi)_{ii}^{-1}$ which is a

nonlinear function of random ξ_i vectors. However the CIs still hold their nominal level of coverage.

- We can estimate parameters via bootstrap if we know distribution of random vector (Y, ξ), but since we do not use the estimated vector (Y, X)