# Notes on Applied Statistics

Zhenbang Wang

This note is partially based on Statistical Concepts and Methods by G.K. Bhattacharyya and R.A. Johnson.

## Contents

# 1 Basic Concepts of Testing Hypothesis

1. A *statistical hypothesis* is a statement about the population. Its plausibility is to be evaluated on the basis of information obtained by sampling from the population.

2. Hypothesis $H$: The proportion of consumers preferring brand A to brand B is 0.4. Hypothesis $H^{'}$ : Above is not true.

3. Difference between mathematical proposition and statistical hypothesis is *uncertainty*

4. The null is the negation of the assertion

5. Type I error: rejection of null when null is true.

6. Type II error: failure to reject null when alternative is true.

7. Power function : $\gamma(\theta) = \mathbf{P}$ [rejects the Null when true value of the parameter is $\theta$]

### 1.1 Selecting among several tests

Level of significance : max $\gamma(p) \leq \alpha$ under p $\in H_0$
Size of the test : max $\gamma(p) = \alpha$ under p $\in H_0$
$p$ - value$(x) = \sup_{\theta \in \Theta_0} P(X \geq x)$
p - value : probability of obtaining a test statistic value as extreme as or more extreme than the observed value under $H_0$

### 1.2 General steps in testing hypothesis

## 2 The Normal Distribution and Random Samples

### 2.1 Random Samples, Statistic, Sampling Distributions

A *random sample* of size n from a population $f(x)$ is a collection of n independent random variables $X_1, X_2, ...$, each having the distribution $f(x)$
A *statistic* is a function of the sample observations.
Every statistic is, itself, a random variable. Its probability distribution is called the sampling distribution of the statistic.

### 2.2 Distribution of the sample mean and the central limit theorem

Mean and standard deviation of $\bar{X}$ : $E(\bar{X}) = \mu$ and $Var(\bar{X}) = \frac{\sigma^2}{n}$
Central limit Theorem: In random sampling from an arbitrary population with mean $\mu$ and standard deviation $\sigma$, the distribution of $\bar{X}$ when n is large is approximately normal, with mean $\mu$ and standard deviation $\sigma/\sqrt{n}$.
In other words, $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is approximately $N(0, 1)$.
e.g. X follows Bernoulli(p). Then $Z = \frac{\hat{p}-p}{\sqrt{p(1-p)/n}}$ is approximately $N(0, 1)$.

### 2.3 Checking the assumption of a normal population

## 3 Inferences about a population

### 3.1 Point Estimation of parameter

Stand error: The standard deviation of the estimator $\hat{\theta}$ is called its standard error and is designated $S.E.(\hat{\theta})$
Point estimator of the mean: $\bar{X}$ $S.E.(\bar{X}) = \sigma/\sqrt{n}$ , estimated $S.E.(\bar{X}) = s/\sqrt{n}$
Point estimator of the binomial Parameter: $\hat{p} = \frac{\sum X}{n}$ $S.E.(\hat{p}) = \sqrt{\frac{pq}{n}}$ and estimated $S.E.(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}$

### 3.2 Estimation by confidence interval

CI: (L,U) is 100(1-$\alpha$)% CI such that $P(L(X_1, X_2, ..., X_n) < \theta < U(X_1, X_2, ..., X_n)) = 1 - \alpha$

Note P[41.1< $\mu$ <44.3] = .95 is wrong.

Interpretation : If we conduct the same experiment independently many times, the confidence interval estimator will cover the true value of $\theta$ approximately $1 - \alpha$ of the time.

Large Sample Confidence Interval for $\mu$ When $\sigma$ is Unknown: $(\bar{X} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{s}{\sqrt{n}})$

Large Sample Confidence Interval for $\sigma^2$: $(\frac{(n-1)S^2}{\mathcal{X}^2_{n-1,\alpha/2}}, \frac{(n-1)S^2}{\mathcal{X}^2_{n-1,1-\alpha/2}})$

Large Sample Confidence Interval for p : $(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}})$

Small Sample Confidence Interval for $\mu$ when $\sigma$ is unknown: $(\bar{X} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}, \bar{X} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}})$

To be $100(1-\alpha)$ sure that the error $|\bar{X} - \mu|$ does not exceed $d$, the required sample size is $n = (\frac{z_{\alpha/2\sigma}}{d})^2$

If n is small and the population is nonnormal,take $\frac{\sigma^2}{\alpha d^2}$ as upper bound.

## 3.3 One Sample binomial proportion

$X_1, X_2, ..., X_n \sim$ bin$(1, p)$ with $X = \sum_{i=1}^{n} X_i$

| | $H_0 : p \leq p_0$ vs $H_1 : p > p_0$ | $H_0 : p \geq p_0$ vs $H_1 : p < p_0$ | $H_0 : p = p_0$ vs $H_1 : p \neq p_0$ |
|---|---|---|---|
| Test statistic | | $X \sim$ bin$(n, p)$ | |
| RR | $\{X \geq c\}$ | $\{X \leq c\}$ | $\{X \leq c_1$ or $X \geq c_2\}$ |
| p-value(x) | $P(X \geq x \mid p_0)$ | $P(X \leq x \mid p_0)$ | $P(\lvert X - np\rvert \geq \lvert x - np\rvert \mid p_0)$ |
| Large Sample version | | | |
| Test statistic | | $X \sim N(np, np(1-p))$ | |
| RR : | $\{X \geq c\}$ | $\{X \leq c\}$ | $\{X \leq c_1$ or $X \geq c_2\}$ |
| p-value(x) | $P(Z \geq \frac{x-np_0}{\sqrt{np_0(1-p_0)}})$ | $P(Z \leq \frac{x-np_0}{\sqrt{np_0(1-p_0)}})$ | $P(\lvert Z\rvert \geq \frac{\lvert x-np_0\rvert}{\sqrt{np_0(1-p_0)}})$ |

## 3.4 One Normal sampling

$X_1, ..., X_n \sim N(\mu, \sigma^2)$ with $\sigma^2$ known

| | $H_0 : \mu \leq \mu_0$ vs $H_1 : \mu > \mu_0$ | $H_0 : \mu \geq \mu_0$ vs $H_1 : \mu < \mu_0$ | $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$ |
|---|---|---|---|
| Test statistic | | $\bar{X}$ or $Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$ | |
| RR | $\{Z \geq c\}$ | $\{Z \leq c\}$ | $\{Z \leq c_1$ or $Z \geq c_2\}$ |
| p-value(x) | $P(Z \geq z \mid \mu_0) = 1 - \phi(z)$ | $P(Z \leq z \mid \mu_0) = \phi(z)$ | $P(\lvert Z\rvert \geq \lvert z\rvert \mid \mu_0) = 2\phi(-\lvert z\rvert)$ |

$X_1, ..., X_n \sim N(\mu, \sigma^2)$ with $\sigma^2$ unknown

| | $H_0 : \mu \leq \mu_0$ vs $H_1 : \mu > \mu_0$ | $H_0 : \mu \geq \mu_0$ vs $H_1 : \mu < \mu_0$ | $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$ |
|---|---|---|---|
| Test statistic | | $T = \frac{\bar{X}-\mu_0}{S/\sqrt{n}}$ | |
| RR | $\{T \geq c\}$ | $\{T \leq c\}$ | $\{T \leq c_1$ or $T \geq c_2\}$ |
| p-value(x) | $P(T \geq t \mid \mu_0) = 1 - F_{t_{n-1}}(t)$ | $P(T \leq t \mid \mu_0) = F_{t_{n-1}}(t)$ | $P(\lvert T\rvert \geq \lvert t\rvert \mid \mu_0) = 2F_{t_{n-1}}(-\lvert t\rvert)$ |

Large Sample Approximation :

Test statistic : $T = \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \sim N(0, 1)$

RR and P - value are the same with $\sigma^2$ known case.

## 3.5 Inference about $\sigma^2$ of a normal population

Test statistic : $\frac{(n-1)S^2}{\sigma_0^2}$

---

$H_0 : \sigma^2 \leq \sigma_0^2$ vs $H_1 : \sigma^2 > \sigma_0^2$

RR : $\frac{(n-1)S^2}{\sigma_0^2} \geq \mathcal{X}^2_{n-1,\alpha}$

p-value(x) : $1 - F_{\mathcal{X}^2_{n-1}}(\frac{(n-1)S^2}{\sigma_0^2})$

---

$H_0 : \sigma^2 \geq \sigma_0^2$ vs $H_1 : \sigma^2 < \sigma_0^2$

RR : $\frac{(n-1)S^2}{\sigma_0^2} \leq \mathcal{X}^2_{n-1,1-\alpha}$

p-value(x) : $F_{\mathcal{X}^2_{n-1}}(\frac{(n-1)S^2}{\sigma_0^2})$

---

$H_0 : \sigma^2 = \sigma_0^2$ vs $H_1 : \sigma^2 \neq \sigma_0^2$

RR : $\frac{(n-1)S^2}{\sigma_0^2} \leq \mathcal{X}^2_{n-1,1-\alpha/2}$ or $\frac{(n-1)S^2}{\sigma_0^2} \geq \mathcal{X}^2_{n-1,\alpha/2}$

P- value is complicated!

Notes : the inference procedures for $\sigma^2$ presented here are extremely sensitive to departures from a normal population!

# 4 Comparing Two treatments

## 4.1 Independent Samples from two populations

Let $X_1, ...X_{n_1} \sim$ i.i.d. as $N(\mu_1, \sigma_1^2)$ and let $Y_1, ..., Y_{n_2} \sim$ i.i.d as $N(\mu_2, \sigma_2^2)$. and $\sigma_1^2 = \sigma_2^2 = \sigma^2$ unknown

Test statistic : $T = \frac{\bar{X}-\bar{Y}-\delta_0}{S_{\text{pooled}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}$

---

$H_0 : \mu_1 - \mu_2 = \delta_0$ vs $H_1 : \mu_1 - \mu_2 \neq \delta_0$

RR : $|T| > t_{n_1+n_2-2,\alpha/2}$

p -value : $P(|T| \geq |t||\mu_1 - \mu_2 = \delta_0) = 2F_{t_{n_1+n_2-2}}(-|t|)$

---

$H_0 : \mu_1 - \mu_2 = \delta_0$ vs $H_1 : \mu_1 - \mu_2 > \delta_0$

RR : $T \geq t_{n_1+n_2-2,\alpha}$

p -value : $P(T \geq t|\mu_1 - \mu_2 = \delta_0) = 1 - F_{t_{n_1+n_2-2}}(t)$

---

$H_0 : \mu_1 - \mu_2 = \delta_0$ vs $H_1 : \mu_1 - \mu_2 < \delta_0$

RR : $T \leq t_{n_1+n_2-2,\alpha}$

p -value : $P(T \leq t|\mu_1 - \mu_2 = \delta_0) = F_{t_{n_1+n_2-2}}(t)$

---

Let $X_1, ...X_{n_1} \sim$ i.i.d. as $N(\mu_1, \sigma_1^2)$ and let $Y_1, ..., Y_{n_2} \sim$ i.i.d as $N(\mu_2, \sigma_2^2)$. and $\sigma_1^2 \neq \sigma_2^2$ known

$H_0 : \mu_1 - \mu_2 = \delta_0$

Test statistic : $Z = \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$

Welch's t test or Behrens-Fisher problem :

Let $X_1, ...X_{n_1} \sim$ i.i.d. as $N(\mu_1, \sigma_1^2)$ and let $Y_1, ..., Y_{n_2} \sim$ i.i.d as $N(\mu_2, \sigma_2^2)$. and $\sigma_1^2 \neq \sigma_2^2$ unknown

$H_0 : \mu_1 - \mu_2 = \delta_0$

Test statistic : $T = \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \sim t_v$ where $v = \frac{(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2})^2}{\frac{1}{n_1-1}(\frac{S_1^2}{n_1})^2+\frac{1}{n_2-1}(\frac{S_2^2}{n_2})^2}$

Large Sample inferences version :

Let $X_1, ...X_{n_1} \sim$ i.i.d. as $N(\mu_1, \sigma_1^2)$ and let $Y_1, ..., Y_{n_2} \sim$ i.i.d as $N(\mu_2, \sigma_2^2)$. and $\sigma_1^2 \neq \sigma_2^2$ unknown

$H_0 : \mu_1 - \mu_2 = \delta_0$

Test statistic : $Z = \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \sim N(0,1)$

## 4.2 Comparing the variances of two normal populations

Let $X_1, ...X_{n_1} \sim$ i.i.d. as $N(\mu_1, \sigma_1^2)$ and let $Y_1, ..., Y_{n_2} \sim$ i.i.d as $N(\mu_2, \sigma_2^2)$.

Test statistic : $F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$

---

$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$ vs $H_1 : \sigma_1^2 > \sigma_2^2$

$RR = \frac{S_1^2}{S_2^2} \geq F_{(n_1-1, n_2-1, \alpha)}$

---

$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$ vs $H_1 : \sigma_1^2 < \sigma_2^2$

$RR = \frac{S_1^2}{S_2^2} \leq F_{(n_1-1, n_2-1, \alpha)}$

---

$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$ vs $H_1 : \sigma_1^2 \neq \sigma_2^2$

$RR = \frac{S_1^2}{S_2^2} \geq F_{(n_1-1, n_2-1, \alpha/2)}$ or $\frac{S_1^2}{S_2^2} \leq F_{(n_1-1, n_2-1, 1-\alpha/2)}$

## 4.3 Comparing Two proportions

$X \sim$ Binomial$(n_1, p_1)$ and $Y \sim$ Binomial$(n_2, p_2)$, $\hat{p_1} = \frac{X}{n_1}$ and $\hat{p_2} = \frac{Y}{n_2}$

Test statistic : $\frac{\hat{p_1}-\hat{p_2}-(p_1-p_2)}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1}+\frac{\hat{p_2}(1-\hat{p_2})}{n_2}}} \sim N(0,1)$

$H_0 : p_1 = p_2$

Large samples version : $Z = \frac{\hat{p_1}-\hat{p_2}-0}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ where $\hat{p} = \frac{X+Y}{n_1+n_2}$

## 4.4 Paired Comparisons

$D_i = X_i - Y_i$ are independent with $N(\delta, \sigma_D^2)$. Let $\bar{D} = \sum_{i=1}^n D_i/n$, $S_D = \sqrt{\sum_{i=1}^n (D_i - \bar{D})^2/(n-1)}$

Test statistic : $T = \frac{\bar{D}-\delta_0}{S_D/\sqrt{n}} \sim t_{n-1}$

---

$H_0 : \delta = \delta_0$ vs $H_1 : \delta > \delta_0$

RR : $\{T \geq c\}$

p - value(x) : $P(T \geq x|\delta_0) = 1 - F_{t_{n-1}}(x)$

---

$H_0 : \delta = \delta_0$ vs $H_1 : \delta < \delta_0$

RR : $\{T \leq c\}$

p - value(x) : $P(T \leq x|\delta_0) = F_{t_{n-1}}(x)$

---

$H_0 : \delta = \delta_0$ vs $H_1 : \delta \neq \delta_0$

RR : $\{|T| \leq c\}$

p - value(x) : $P(|T| \leq |x||\delta_0) = 2F_{t_{n-1}}(-|x|)$

A 100$(1-\alpha)$ Confidence Interval for $\delta$ is given by : $\bar{D} \mp t_{n-1,\alpha/2} S_D/\sqrt{n}$

# 5 Design of Experiments and Analysis of Variance

## 5.1 Comparison of several treatments

Data Structure :

| | Treatment 1 | Treatment 2 | $\cdots$ | Treatment K |
|---|---|---|---|---|
| | $y_{11}$ | $y_{12}$ | $\cdots$ | $y_{1K}$ |
| | $y_{21}$ | $y_{12}$ | $\cdots$ | $y_{2K}$ |
| | $\vdots$ | $\vdots$ | $\cdots$ | $\vdots$ |
| | $y_{n_1 1}$ | $y_{n_2 2}$ | $\cdots$ | $y_{n_K K}$ |
| Means | $\bar{y}_1$ | $\bar{y}_2$ | $\cdots$ | $\bar{y}_K$ |

Decomposition of $y_{ij} = \bar{y} + (\bar{y}_j - \bar{y}) + (y_{ij} - \bar{y}_j)$

| Source | SS | DF | MS | F |
|---|---|---|---|---|
| Treatments | SST = $\sum_{j=1}^{K} n_j(\bar{y}_j - \bar{y})^2$ | K - 1 | $\frac{SST}{K-1}$ | $\frac{MS(T)}{MS(E)}$ |
| Error | SSE = $\sum_{j=1}^{K} \sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2$ | $\sum_{j=1}^{K} n_j - K$ | $\frac{SSE}{\sum_{j=1}^{K} n_j - K}$ | |
| Total | $\sum_{j=1}^{K} \sum_{i=1}^{n_j}(y_{ij} - \bar{y})^2$ | $\sum_{j=1}^{K} n_j - 1$ | | |

## 5.2 Population model and inferences

$Y_{ij} = \mu + \beta_j + e_{ij}$, $j = 1, ..., K$, $i = 1, ..., n_j$
where $\mu$ is overall mean and $\beta_j$ is the jth treatment effect, $\sum_{j=1}^{K} \beta_j = 0$, and $e_{ij}$ are i.i.d. $N(0, \sigma^2)$.

The likelihood ratio test or F test of the null hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = ... = \beta_K = 0$ vs $H_1$ : some of the $\beta_j$ values differ from zero is by using $F$ from above. Under $H_0$, $F = \frac{MS(T)}{MS(E)} \sim F_{K-1,N-K}$.
RR : $F > F_{K-1,N-K,\alpha}$
p-value : $P(F >$ observed value)

Confidence interval for a single difference $(\mu_j - \mu_{j'})$:
$\bar{y}_j - \bar{y}_{j'} \pm t_{N-K,\alpha/2}\sqrt{MSE}\sqrt{\frac{1}{n_j} + \frac{1}{n_{j'}}}$

Multiple-t Confidence Intervals (Bonferroni Intervals)
A set of $100(1-\alpha)$ simultaneous confidence intervals for $m = \binom{K}{2}$ number of pairwise differences $(\mu_j - \mu_{j'})$ is given by

$$\bar{y}_j - \bar{y}_{j'} \pm t_{N-K,\alpha/2m}\sqrt{MSE}\sqrt{\frac{1}{n_j} + \frac{1}{n_{j'}}}$$

## 5.3 One way ANOVA

### 5.3.1 Fixed Effect

$y_{ij} = \mu + \beta_i + e_{ij}$, $i = 1, ..., K$, $j = 1, ..., n_i$
where $\mu$ is overall mean, $\beta_i$ is the ith treatment effect, $e_{ij}$ are i.i.d.$N(0, \sigma_e^2)$

### 5.3.2 Random effect

$y_{ij} = \mu + \alpha_i + e_{ij}, i = 1, ..., K, j = 1, ..., n_i$
where $\alpha_i$ i.i.d. $N(0, \sigma_a^2)$ and independent of $e_{ij}$, $e_{ij}$ are i.i.d.$N(0, \sigma_e^2)$

The F test of the null hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = ... = \beta_K = 0$
vs $H_1$ : some of the $\beta_j$ values differ from zero is by using $F$ from above. $H_0 : \sigma_a^2 = 0$ vs $H_1 : \sigma_a^2 > 0$. Under $H_0$, $F = \frac{MS(T)}{MS(E)} \sim F_{K-1,N-K}$.
RR : $F > F_{K-1,N-K,\alpha}$
p-value : $P(F >$ observed value)

## 5.4 Two - Sample Median Test

$H_0$ :Both population medians are the same
$H_1$ :Population medians differ

## 5.5 Randomized Block Experiments

Data Structure :

| | Block 1 | Block 2 | $\cdots$ | Block b | Treatment means |
|---|---|---|---|---|---|
| Treatment 1 | $y_{11}$ | $y_{12}$ | $\cdots$ | $y_{1b}$ | $\bar{y}_{1\cdot}$ |
| Treatment 2 | $y_{21}$ | $y_{11}$ | $\cdots$ | $y_{2b}$ | $\bar{y}_{2\cdot}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\cdots$ | $\vdots$ | $\vdots$ |
| Treatment K | $y_{K1}$ | $y_{K2}$ | $\cdots$ | $y_{Kb}$ | $\bar{y}_{K\cdot}$ |
| Block means | $y_{\cdot 1}$ | $y_{\cdot 2}$ | $\cdots$ | $y_{\cdot b}$ | $\bar{y}_{\cdot\cdot}$ |

Decomposition of $y_{ij} = (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) + (\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}) + (y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot}) + \bar{y}_{\cdot\cdot}$

| Source | SS | DF | MS | F |
|---|---|---|---|---|
| Treatments | SST = $b \sum_{i=1}^{K} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$ | K - 1 | $\frac{SST}{K-1}$ | $\frac{\text{MST}}{\text{MSE}}$ |
| Blocks | SSB = $K \sum_{j=1}^{b} (\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2$ | b - 1 | $\frac{SSB}{b-1}$ | $\frac{\text{MSB}}{\text{MSE}}$ |
| Error | SSE = $\sum_{i=1}^{K} \sum_{j=1}^{b} (y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot})^2$ | $(K-1)(b-1)$ | $\frac{SSE}{(K-1)(b-1)}$ | |
| Total | $\sum_{j=1}^{K} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{\cdot\cdot})^2$ | $bK - 1$ | | |

Population model :
$Y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$ , $i = 1, ..., K$, $j = 1, ..., b$ where $\sum_{i=1}^{K} \alpha_i = 0$, $\sum_{j=1}^{b} \beta_j = 0$ and $e_{ij} \sim N(0, \sigma^2)$.
Testing :
Reject $H_0$: $\alpha_1 = \alpha_2 = \cdots = \alpha_K = 0$ (no treatment differences) if $\frac{\text{MST}}{\text{MSE}} > F_{K-1,(K-1)(b-1),\alpha}$
Reject $H_0 : \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_b = 0$ (no block differences) if $\frac{\text{MSB}}{\text{MSE}} > F_{b-1,(K-1)(b-1),\alpha}$
Confidence interval :
A $100(1-\alpha)$ confidence interval for $(\beta_i - \beta_{i'})$ is given by $(\bar{y}_{i\cdot\cdot} - \bar{y}_{i'\cdot}) \pm t_{(b-1)(K-1),\alpha/2} \sqrt{\text{MSE} \cdot 2/b}$

## 5.6 Factorial Experiment(Interaction)

Data Structure : Suppose we have $r(r > 1)$ replicates, i.e., we repeat the experiment r times using r sets of pq experimental units. Factor A has p levels and Factor B has q levels

| | B 1 | B 2 | $\cdots$ | B q |
|---|---|---|---|---|
| A 1 | $y_{11}$ | $y_{12}$ | $\cdots$ | $y_{1q}$ |
| A 2 | $y_{21}$ | $y_{11}$ | $\cdots$ | $y_{2q}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\cdots$ | $\vdots$ |
| A p | $y_{p1}$ | $y_{p2}$ | $\cdots$ | $y_{pq}$ |

Decomposition of $y_{ijk} = \bar{y}_{\cdots} + (\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots}) + (\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots}) + (y_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots}) + (y_{ijk} - \bar{y}_{ij\cdot})$

| Source | SS | DF | MS | F |
|---|---|---|---|---|
| Factor A | SSA = $qr\sum_{i=1}^{p}(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots})^2$ | p - 1 | $\frac{SSA}{p-1}$ | $\frac{\text{MSA}}{\text{MSE}}$ |
| Factor B | SSB = $pr\sum_{j=1}^{q}(\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots})^2$ | q - 1 | $\frac{SSB}{q-1}$ | $\frac{\text{MSB}}{\text{MSE}}$ |
| Interaction A×B | SSAB = $r\sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots})^2$ | $(p-1)(q-1)$ | $\frac{SSAB}{(q-1)(p-1)}$ | $\frac{\text{MSAB}}{\text{MSE}}$ |
| Error | SSE = $\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r}(y_{ijk} - \bar{y}_{ij\cdot})^2$ | $pq(r-1)$ | $\frac{SSE}{pq(r-1)}$ | |
| Total | $\sum_{k=1}^{r}\sum_{j=1}^{q}\sum_{i=1}^{p}(y_{ijk} - \bar{y}_{\cdots})^2$ | $pqr-1$ | | |

Population model :

$Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$ , $i = 1, ..., p, j = 1, ..., q, k = 1, ..., n_{ij} = r.$

Testing :

Reject $H_0$: $\alpha_1 = \alpha_2 = \cdots = \alpha_p = 0$ if $\frac{\text{MSA}}{\text{MSE}} > F_{p-1,pq(r-1),\alpha}$

Reject $H_0 : \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_q = 0$ if $\frac{\text{MSB}}{\text{MSE}} > F_{q-1,pq(r-1),\alpha}$

Reject $H_0 : \gamma_1 = \gamma_2 = \gamma_3 = \cdots = \gamma_r = 0$ if $\frac{\text{MSAB}}{\text{MSE}} > F_{(p-1)(q-1),pq(r-1),\alpha}$

### 5.7 Two way ANOVA

# 6 Analysis of Categorized Data

## 6.1 The multinomial model

Structure of Multinomial Data:

| Cells | 1 | 2 | $\cdots$ | K | Total |
|---|---|---|---|---|---|
| Probabilities | $p_1$ | $p_2$ | $\cdots$ | $p_K$ | 1 |
| Frequencies in n trials | $n_1$ | $n_2$ | $\cdots$ | $n_K$ | n |

## 6.2 Pearson's Test for Goodness of fit

Case A : Cell Probabilities Completely Specified by $H_0$

Null Hypothesis : $H_0 : p_1 = p_{10}, \cdots , p_k = p_{k0}$

Test statistic : $\mathcal{X}^2 = \sum_{i=1}^{k} \frac{(n_i - np_{i0})^2}{np_{i0}} = \sum_{\text{cell } i} \frac{(O_i - E_i)^2}{E_i}$

The $\mathcal{X}^2$ statistic is approximately $\mathcal{X}^2_{k-1}$ distributed for large n under the null.

RR = $X^2 \geq \mathcal{X}^2_{k-1,\alpha}$ and p - value = $P(\mathcal{X}^2_{k-1} \geq \text{ observed} X^2)$

Large sample approximation if all expected cell counts $\geq 5$

Case B : Cell Probabilities Not Completely Specified by $H_0$

First estimate the unknown parameter under the null assuming a parametric model.

Next calculate the expected cell counts under the null using the parameter value obtained in the first step.

$\mathcal{X}^2 = \sum_{\text{cell } i} \frac{(O_i - E_i)^2}{E_i}$ with d.f. = number of cells - 1 - (number of estimated parameters)

## 6.3 Contingency Tables

An r $\times$ c Contingency Table – Data Structure:

| | $B_1$ | $B_2$ | $\cdots$ | $B_c$ | Row Total |
|---|---|---|---|---|---|
| $A_1$ | $n_{11}$ | $n_{12}$ | $\cdots$ | $n_{1c}$ | $n_{10}$ |
| $A_2$ | $n_{21}$ | $n_{22}$ | $\cdots$ | $n_{2c}$ | $n_{20}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\cdots$ | $\vdots$ | $\vdots$ |
| $A_r$ | $n_{r1}$ | $n_{r2}$ | $\cdots$ | $n_{rc}$ | $n_{r0}$ |
| Column Total | $n_{01}$ | $n_{02}$ | $\cdots$ | $n_{0c}$ | n |

| | $B_1$ | $B_2$ | $\cdots$ | $B_c$ | Row Total |
|---|---|---|---|---|---|
| $A_1$ | $p_{11}$ | $p_{12}$ | $\cdots$ | $p_{1c}$ | $p_{10}$ |
| $A_2$ | $p_{21}$ | $p_{22}$ | $\cdots$ | $p_{2c}$ | $p_{20}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\cdots$ | $\vdots$ | $\vdots$ |
| $A_r$ | $p_{r1}$ | $p_{r2}$ | $\cdots$ | $p_{rc}$ | $p_{r0}$ |
| Column Total | $p_{01}$ | $p_{02}$ | $\cdots$ | $p_{0c}$ | 1 |

Null hypothesis : the A and B classification are independent
$H_0 : p_{ij} = p_{i0}p_{0j}$ for cells $(i, j)$
Under $H_0$, $E(n_{ij}) = np_{i0}p_{0j}$
Estimators of $p_{i0}$ and $p_{0j}$ : $\hat{p_{i0}} = \frac{n_{i0}}{n}$, $\hat{p_{0j}} = \frac{n_{0j}}{n}$
Test statistic : $X^2 = \sum_{\text{all cells}} \frac{(n_{ij}-E_{ij})^2}{E_{ij}}$ where $E_{ij} = \frac{n_{i0}n_{0j}}{n}$
The distribution of $X^2$ under $H_0$ can be approximated by $X^2_{(r-1)(c-1)}$ for large sample (all expected cell counts $\geq 5$ )

### 6.3.1 Measures of Association in a Contingency Table

$q = \min(r, c)$. Large values imply strong association :
Cramer's contingency coefficient :
$Q_1 = \frac{X^2}{n(q-1)}$ , $0 \leq Q_1 \leq 1$
Pearson's coefficient of mean square contingency
$Q_2 = \sqrt{\frac{X^2}{n+X^2}}$, $0 \leq Q_2 \leq \sqrt{\frac{q-1}{q}}$

### 6.3.2 Contingency tables with one margin fixed(Test of Homogeneity)

An r $\times$ c Contingency Table – Data Structure:

| | $B_1$ | $B_2$ | $\cdots$ | $B_c$ | Row Total |
|---|---|---|---|---|---|
| $A_1$ | $n_{11}$ | $n_{12}$ | $\cdots$ | $n_{1c}$ | $n_{10}$ |
| $A_2$ | $n_{21}$ | $n_{22}$ | $\cdots$ | $n_{2c}$ | $n_{20}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\cdots$ | $\vdots$ | $\vdots$ |
| $A_r$ | $n_{r1}$ | $n_{r2}$ | $\cdots$ | $n_{rc}$ | $n_{r0}$ |
| Column Total | $n_{01}$ | $n_{02}$ | $\cdots$ | $n_{0c}$ | n |

| | $B_1$ | $B_2$ | $\cdots$ | $B_c$ | Row Total |
|---|---|---|---|---|---|
| $A_1$ | $w_{11}$ | $w_{12}$ | $\cdots$ | $w_{1c}$ | 1 |
| $A_2$ | $w_{21}$ | $w_{22}$ | $\cdots$ | $w_{2c}$ | 1 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\cdots$ | $\vdots$ | $\vdots$ |
| $A_r$ | $w_{r1}$ | $w_{r2}$ | $\cdots$ | $w_{rc}$ | 1 |

Null hypothesis of homogeneity : $w_{1j} = w_{2j} = \cdots = w_{rj}$ for every $j = 1, ..., c$
The estimated probability is $\hat{w_{1j}} = \hat{w_{2j}} = \cdots = \hat{w_{rj}} = \frac{n_{0j}}{n}$ and the expected frequency in the $(i, j)$th cell is $E_{ij} = n_{i0}\hat{w}_{ij} = \frac{n_{i0}n_{0j}}{n}$
The test statistic is given by $X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-E_{ij})^2}{E_{ij}} \sim \mathcal{X}^2_{(r-1)(c-1)}$

#### 6.3.3 2×2 Contingency Table

$H_0 : p_1 = p_2$ vs $H_1 : p_1 \neq p_2$
Pearson's $\mathcal{X}^2$ test, provided expected cell counts $\geq 5$
$Z$ test by the normal(large - sample) approximations $Z = \frac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{(1/n_1)+(1/n_2)}} \sim N(0, 1)$.
$H_0 : p_1 = p_2$ vs $H_1 : p_1 > p_2$ or $H_1 : p_1 < p_2$
Pearson's $\mathcal{X}^2$ test not appropriate.
Z test by the normal approximations.

#### 6.3.4 I × J Contingency Tables

#### 6.3.5 Fisher's exact test

#### 6.3.6 Ordinal Tests

# 7 Nonparametric Inference

## 7.1 Paired Comparisons

### 7.1.1 The Sign Test

$H_0 : \eta = \eta_0$ [ i.e. $P(y < \eta_0) = P(y > \eta_0) = \frac{1}{2}$, Note this is no ties version.]
The test statistic : $S = \sum_{i=1}^{n} I\{y_i > \eta_0\}$, under the null, $S \sim$ Binomial$(n', p = \frac{1}{2})$ where $n' = n -$ (number of sample equals to $\eta_0$) is called effective sample size.

---

$H_1 : \eta > \eta_0$
$RR = \{S \geq c\}$ where $c$ satisfies $P(S \geq c) \leq \alpha$
p-value$(s) = P(S \geq s)$

---

$H_1 : \eta < \eta_0$
$RR = \{S \leq c\}$ where $c$ satisfies $P(S \leq c) \leq \alpha$
p-value$(s) = P(S \leq s)$

---

$H_1 : \eta \neq \eta_0$
$RR = \{S \leq c_1$ or $S \geq c_2\}$ where $c$ satisfies $P(S \leq c_1) + P(S \geq c_2) \leq \alpha$

$$\text{p-value}(s) = \begin{cases} P(S \leq s) + P(S \geq n' - s) & s \leq \frac{n'}{2} \\ P(S \leq n' - s) + P(S \geq s) & s \geq \frac{n'}{2} \end{cases}$$

---

Note that for $n^{'} > 25$, we can use $\frac{S-\frac{n^{'}}{2}}{\frac{1}{2}\sqrt{n^{'}}} \sim N(0,1)$ as test statistic.

### 7.1.2 The Wilcoxon Signed Rank Test

The null hypothesis : the underlying $cdf$ is symmetric about a specified value $\eta_0$.
Steps in the signed-rank test:

1. Discard values of $X_i = \eta_0$
2. Let $Y_i = X_i - \eta_0$, let $r_i$ be the rank of $|Y_i|$ if there is a tie then average ranks for tied values
3. Define $T_+ = \sum_{i=1}^{n^{'}} r_i I\{Y_i > 0\} = \sum_{i=1}^{n^{'}} \sum_{j=1}^{i} I\{Y_i + Y_j > 0\}$

---

$H_1 : \eta > \eta_0$
$RR = \{T^+ \geq c\}$ where $c$ satisfies $P(T^+ \geq c|H_0) \leq \alpha$
p-value$(t^+) = P(T^+ \geq t^+)$

---

$H_1 : \eta < \eta_0$
$RR = \{T^+ \leq c\}$ where $c$ satisfies $P(T^+ \leq c|H_0) \leq \alpha$
p-value$(t^+) = P(T^+ \leq t^+)$

---

$H_1 : \eta \neq \eta_0$
$RR = \{T^+ \leq c_1$ or $T^+ \geq c_2\}$ where $c$ satisfies $P(T^+ \leq c_1) + P(T^+ \geq c_2) \leq \alpha$

$$\text{p-value}(s) = \begin{cases} P(T^+ \leq t^+|H_0) + P(T^+ \geq \frac{n^{'}(n^{'}+1)}{2} - t^+|H_0) & t^+ \leq \frac{n^{'}(n^{'}+1)}{4} \\ P(T^+ \geq t^+|H_0) + P(T^+ \leq \frac{n^{'}(n^{'}+1)}{2} - t^+|H_0) & t^+ > \frac{n^{'}(n^{'}+1)}{4} \end{cases}$$

---

When $n^{'} > 25$, we can use $\frac{T_+ - \frac{n^{'}(n^{'}+1)}{4}}{\sqrt{\frac{n^{'}(n^{'}+1)(2n^{'}+1)}{24}}} \sim N(0,1)$ as test statistic.

## 7.2 The Wilcoxon Rank-Sum test for comparing two treatments

$H_0$ : The two population distributions are identical.[i.e. $F_{X_A} = F_{X_B}$]
One side alternative :
$H_1$: The distribution of population A is shifted to the right/left of the distribution of population B.
Two sided alternative :
$H_1$: The distribution of population A is different from the distribution of population B.

Test statistic : $W_A = \sum_{i=1}^{n_A} R(X_{1i})$ where $R(X_{1i})$ is the rank of $X_{1i}$ in the pooled sample , $W_A$ is the rank sum for treatment A and $W_A$ is symmetric about $n_A(n_A + n_B + 1)/2$ under $H_0$
Note : We could use $W_B = \sum_{i=1}^{n_B} R(X_{2i})$ where $R(X_{2i})$ is the rank of $X_{2i}$ in the pooled sample, $W_B$ is the rank sum for treatment B and $W_B$ is symmetric about $n_B(n_A + n_B + 1)/2$ under $H_0$

$W_A + W_B = \frac{(n_A+n_B)(n_A+n_B+1)}{2}$ which is a constant.
Let $W_s$ = sum of ranks of the smaller sample. [i.e. Determine whether $W_s = W_A$ or $W_s = W_B$]

For $H_1$: Population A is shifted to the right of population B; set the rejection region of the form $W_s \geq c$.

For $H_1$: Population A is shifted to the left of population B; set the rejection region of the form $W_s \leq c$.

For $H_1$: Populations are different; set the rejection region of the form $W_s \leq c_1$ or $W_s \geq c_2$.

Large Sample Approximation :
Under $H_0$ mean of $W_A = \frac{n_A(n_A+n_B+1)}{2}$, variance of $W_A = \frac{n_A n_B(n_A+n_B+1)}{12}$
Test statistic : $Z = \frac{W_A - \frac{n_A(n_A+n_B+1)}{2}}{\sqrt{\frac{n_A n_B(n_A+n_B+1)}{12}}} \sim N(0,1)$

## 7.3 The Kruskal-Wallis Test

$H_0$ : All K continuous population distributions are identical
$H_1$ : Not all K distributions are identical

Notes: 1. When $K = 2$, Kruskal-Wallis Test and Wilcoxon Rank sum test are the same.
2. Kruskal-Wallis setup is akin to conducting an ANOVA F test on ranks instead of $y_{ij}$.

Procedures: 1. Get the rank table

| | Treatment 1 | Treatment 2 | $\cdots$ | Treatment K |
|---|---|---|---|---|
| | $R_{11}$ | $R_{12}$ | $\cdots$ | $R_{1K}$ |
| | $R_{21}$ | $R_{11}$ | $\cdots$ | $R_{2K}$ |
| | $\vdots$ | $\vdots$ | $\cdots$ | $\vdots$ |
| | $R_{n_1 1}$ | $R_{n_2 1}$ | $\cdots$ | $R_{n_K K}$ |
| Rank sum | $W_1$ | $W_2$ | $\cdots$ | $W_K$ |
| Average Rank | $\bar{R}_1$ | $\bar{R}_2$ | $\cdots$ | $\bar{R}_K$ |

2 . The pooled-sample average rank is $\bar{R} = \frac{1+2+\cdots+N}{N} = \frac{N+1}{2}$.
3. Under $H_0$ , the sample average ranks are all close to the pooled average $\bar{R}$.
4. The Kruskal-Wallis statistic $H = \frac{12}{N(N+1)} \sum_{i=1}^{K} n_i(\bar{R}_i - \frac{N+1}{2})^2$ or $H = \frac{12}{N(N+1)}[\frac{W_1^2}{n_1} + \frac{W_2^2}{n_2} + \cdots + \frac{W_K^2}{n_K}] - 3(N+1)$.
5. Large values of H support $H_1$.
6. Approximately, $H \sim \mathcal{X}_{K-1}^2$ under $H_0$ for large samples.
7. p - value = $P(\mathcal{X}_{k-1}^2 \geq$ observed H value)

## 7.4 Friedman's rank test

The Friedman test is a non-parametric test for analyzing two-way models without interaction.
Extension of the sign test with more than two treatments.
Model : $Y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$
$H_0$ : The treatment effects of factor A have identical effects
$H_1$ : At least one treatment of factor A is different from at least one other treatment

$H_0$ : The treatment effects of factor B have identical effects
$H_1$ : At least one treatment of factor B is different from at least one other treatment

| | B 1 | B 2 | $\cdots$ | B q |
|---|---|---|---|---|
| A 1 | $y_{11}$ | $y_{12}$ | $\cdots$ | $y_{1q}$ |
| A 2 | $y_{21}$ | $y_{11}$ | $\cdots$ | $y_{2q}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\cdots$ | $\vdots$ |
| A p | $y_{p1}$ | $y_{p2}$ | $\cdots$ | $y_{pq}$ |

Test statistic : $Q_A = \frac{12q}{p(p+1)} \sum_{i=1}^{p} (\bar{R}_{i\cdot} - \frac{p+1}{2})^2 \sim \mathcal{X}_{p-1}^2$ where $\bar{R}_{i\cdot} = \frac{1}{q} \sum_{j=1}^{q} R_{ij}$

Test statistic : $Q_B = \frac{12p}{q(q+1)} \sum_{i=1}^{q} (\bar{R}_{\cdot j} - \frac{q+1}{2})^2 \sim \mathcal{X}^2_{q-1}$ where $\bar{R}_{\cdot j} = \frac{1}{p} \sum_{j=1}^{p} R_{ij}$
Large Value of $Q_A$ and $Q_B$ support $H_1$.

# 8 Simple Linear Relation

Note : Be careful with population variance , sample variance of $Y$ [i.e sample variance of Y $= S_y^2/(n-1)$]

## 8.1 Correlation coefficient

Sample Correlation coefficient : $r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{[\sum_{i=1}^{n}(X_i-\bar{X})^2][\sum_{i=1}^{n}(Y_i-\bar{Y})^2]}} = \frac{S_{xy}}{\sqrt{S_x^2 S_y^2}}$

## 8.2 Simple linear regression

Statistical Model : $Y_i = \alpha + \beta x_i + e_i$ , $i = 1, ..., n$
$x_1, ..., x_n$ are the set values of the independent variable x.
$e_1, ..., e_n$ are the unknown random error, which we assume are i.i.d $N(0, \sigma^2)$.
The intercept $\alpha$ and slope $\beta$ are unknown.
$E(Y|x) = \alpha + \beta x$, i.e. the mean response changes linearly with x.

The Principle of Least Squares :
Least squares regression (fitted) line : $\hat{y} = \hat{\alpha} + \hat{\beta} x$
Residual or Error Sum of Squares : $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta} x)^2$
$\hat{\alpha}$ and $\hat{\beta}$ are selected to give minimum SSE

Formulas for Least Squares Estimates :
$\hat{\beta} = \frac{S_{xy}}{S_x^2}$ and $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ where $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ and $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$, $S_x^2 = \sum(x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2$, $S_y^2 = \sum(y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2$, $S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y}$

ANOVA-Type Formulas :
$SS_{\text{Total}} = SS_{\text{Regn}} + SSE$, i.e. total variation = variation due to regression + residual variation.
where $SS_{\text{Total}} = \sum(y_i - \bar{y})^2 = S_y^2$, $SS_{\text{Regn}} = \sum(\hat{y} - \bar{y})^2 = \hat{\beta}^2 S_x^2$, $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = S_y^2 - \hat{\beta}^2 S_x^2$

$R^2 = \frac{SS_{\text{Regn}}}{SS_{\text{Total}}}$, $R^2$ represents the proportion of the y variability explained by the linear relation with x.

Other inference on $\alpha$ and $\beta$ :
$\hat{\alpha} \sim N(\alpha, \sigma^2[\frac{1}{n} + \frac{\bar{x}^2}{S_x^2}])$
$\hat{\beta} \sim N(\beta, \frac{\sigma^2}{S_x^2})$
$s^2 = SSE/(n-2)$ is an unbiased estimator of $\sigma^2$
$(n-2)s^2/\sigma^2 \sim \mathcal{X}^2_{n-2}$ and is independent of $\hat{\alpha}$ and $\hat{\beta}$
Standard error estimate of $\hat{\alpha} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_x^2}}$
Standard error estimate of $\hat{\beta} = \frac{s}{S_x}$

$\frac{(\hat{\beta}-\beta)}{s/S_x} \sim t_{n-2}$
$\frac{(\hat{\alpha}-\alpha)}{s\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_x^2}}} \sim t_{n-2}$

Inference Concerning the Slope $\beta$
Hypothesis Testing :
Test $H_0 : \beta = \beta_0$ vs $H_1 : \beta \neq \beta_0$ or $H_1 : \beta > \beta_0$ or $H_1 : \beta < \beta_0$
Test statistic : $T = \frac{\hat{\beta}-\beta_0}{s/S_x}$
Under $H_0, T \sim t_{n-2}$
Confidence Interval Estimation
$100(1-\alpha)$ CI for $\beta : \hat{\beta} \pm t_{n-2,\alpha/2}\frac{s}{S_X}$

Hypothesis Testing :
Test $H_0 : \alpha = \alpha_0$ vs $H_1 : \alpha \neq \alpha_0$ or $H_1 : \alpha > \alpha_0$ or $H_1 : \alpha < \alpha_0$
Test statistic : $T = \frac{\hat{\alpha}-\alpha_0}{s\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_x^2}}}$
Under $H_0, T \sim t_{n-2}$
Confidence Interval Estimation
$100(1-\alpha)$ CI for $\alpha : \hat{\alpha} \pm t_{n-2,\alpha/2}s\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_x^2}}$

Prediction Interval of the Mean Response for a Specified $x^*$ Value:
$100(1-\alpha)$ CI for $E(Y|x^*) : \hat{\alpha} + \hat{\beta}x^* \pm t_{n-2,\alpha/2}s\sqrt{\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_x^2}}$

Prediction Interval of a Single Response for a Specified $x^*$ Value:
$100(1-\alpha)$ CI : $\hat{\alpha} + \hat{\beta}x^* \pm t_{n-2,\alpha/2}s\sqrt{1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_x^2}}$

# 9 Logistic and Poisson Regression Models

## 9.1 Logistic Regression

$\ln\frac{\pi}{1-\pi} = \mathbf{x}'\beta$

### 9.1.1 Estimating the Parameters in a Logistic Regression Model

### 9.1.2 Interpretation of the Parameters in a Logistic Regression Model

Assume that odds $\frac{\pi}{1-\pi} = e^{\hat{\eta}} = e^{b_0+b_1x_1}$, then estimated odds ratio is $O_R = \frac{\text{odds}_{x_i+1}}{\text{odds}_{x_i}} = e^{b_1}$.

The estimated odds ratio can be interpreted as the estimated increase in the odds of success associated with a one - unit change in the value of the predictor variable.

If $b_1$ is positive, this implies that every additional $(x_1)$ increase the odds of success by $e^{b_1} - 1$ percent.

If $b_1$ is negative, this implies that every additional $(x_1)$ reduce the odds of failure by $1 - e^{b_1}$ percent.

### 9.1.3 Statistical Inference on Model Parameters

Likelihood ratio tests :

Test Goodness of Fit with Deviance :

Test Hypothesis on Subsets of Parameters Using Device :

Test on Individual Model Coefficients :

Lack of Fit Tests in Logistic Regression :

Diagnostic Checking in Logistic Regression :

## 9.2 Poisson Regression

We assume that the response variable $y_i$ is a count, such that the observation $y_i = 0, 1, 2, ...$ $f(y_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}$, $y_i = 0, 1, 2, ...$
Identity Link : $g(\mu_i) = \mu_i = x_i^{'}\beta$
Log Link : $g(\mu_i) = \ln(\mu_i) = x_i^{'}\beta \Rightarrow \mu_i = \exp\{x_i^{'}\beta\}$

Interpretation of $\beta$ : the additive change in the log mean count for each 1-unit increase in x.
Interpretation of $e^{\beta}$ : the multiplicative factor by which the mean count changes for each 1-unit increase in x.