VEE Mathematical Statistics - Formula Sheet

Sample Statistics

Sample Mean: $\bar{X} = \sum_{i=1}^{n} \frac{1}{n} x_i$

Sample Variance, μ known:

$S^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$

Sample Variance, μ unknown:

$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$

Sample statistic following a standard normal distribution: $\dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

Sample statistic following a T-distribution:

$\dfrac{\bar{X} - \mu}{s/\sqrt{n}} \sim T_{n-1} \quad for X \sim N(\mu, \sigma^2), n \leq 30$

Sample statistic following a chi square dist.:

$\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \quad for X \sim N(\mu, \sigma^2)$

Sample statistic following an F-distribution:

$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1, n_2-1} \quad for X_1, X_2 \sim N(\mu, \sigma^2)$

F-distribution critical values:

$f_{k_1,k_2,1-\alpha} = 1/f_{k_2,k_1,\alpha}$

Likelihood

$L(\theta) =$ **Likelihood function**

$L(\theta|X_1, X_2, ..., X_n) = f(X_1, X_2, ..., X_n|\theta)$

$= \prod_{i=1}^{n} f(X_i|\theta)$

$\ell(\theta) = \ln L(\theta) =$ loglikelihood function

$I(\theta) =$ **Fischer Information**

$I(\theta) = E\left[\left[\frac{d\, ln(f(x|\theta))}{d\theta}\right]^2\right]$

$= -E\left[\left[\frac{d^2\, ln(f(x|\theta))}{d^2\theta}\right]\right]$

for a sample of size $n, I_n(\theta) = nI(\theta)$

Cramér-Rao Inequality

$Var(\hat{\theta}) \geq \frac{1 + \frac{d}{d\theta} Bias(\hat{\theta})^2}{nI(\theta)}$

if $\hat{\theta}$ is unbiased, $Var(\hat{\theta}) \geq \frac{1}{nI(\theta)}$

Point Estimates

$\theta =$ Parameter to estimate

$\hat{\theta} =$ Estimate of θ

$\text{bias}_{\hat{\theta}}(\theta) = \text{E}[\hat{\theta}] - \theta$

$\text{Var}[\hat{\theta}] = E[\hat{\theta} - E(\hat{\theta})^2] = \text{E}[\hat{\theta}^2] - \text{E}[\hat{\theta}]^2$

Mean Square Error

$\text{MSE}_{\hat{\theta}}(\theta) = \text{E}[(\hat{\theta} - \theta)^2]$

$\text{MSE}_{\hat{\theta}}(\theta) = \text{Var}[\hat{\theta}] + \left(\text{bias}_{\hat{\theta}}(\theta)\right)^2$

Efficiency: $e(\hat{\theta}) = \dfrac{1/nI(\theta)}{\text{Var}(\hat{\theta})}$

Minimum **V**ariance **U**nbiased **E**stimator

θ is an MVUE if $\text{bias}_{\hat{\theta}}(\theta) = 0$ AND

for all other unbiased $\hat{\theta}, '$ $\text{MSE}_{\hat{\theta}} \leq \text{MSE}_{\hat{\theta}'}$

Consistency

$\hat{\theta}$ is a consistent estimator of θ if

$P[|\hat{\theta} - \theta| > \epsilon] \longrightarrow 0$ as $n \longrightarrow \infty$

practically, if $MSE(\hat{\theta}) \to 0$ as $n \to \infty$

Percentile Matching

$F(\pi_p) = p = p \times 100\%$, the $100p^{\text{th}}$ perecentile

Smoothed emp. per. $\hat{\pi}_{i/(n+1)} = i^{\text{th}}$ obs.

Percentile Matching: Set $\hat{\theta}$ so that $\pi_p = \hat{\pi}_p$

Method of Moments

One Parameter: $E[X] = \bar{X}$

More than one: $E[X^k] = \frac{1}{n} \sum X_i^K$

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

Solve system of equations for parameters

MLE

Procedure:

1. Write $L(\lambda; \mathbb{X})$
2. Take the natural log
3. Compute $\ell'(\lambda; \mathbb{X}) = \frac{d}{d\lambda} \ell(\lambda; \mathbb{X})$
4. Set equal to zero and solve for λ.

MLE = MoM

Exponential (θ)

Gamma (θ when α is known)

Poisson (λ)

Binomial (p when n is known)

Geometric (β)

Neg. Binomial (β when r is known)

Normal (μ, σ^2)

Confidence Intervals

CI on μ, σ^2known, or n large:

Two-sided $100(1-\alpha)\%$CI:

$\bar{x} - z_{\alpha/2} \frac{\sigma ors}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{\sigma ors}{\sqrt{n}}$

Upper One-sided $100(1-\alpha)\%$CI:

$\mu \leq \bar{x} + z_\alpha \frac{\sigma ors}{\sqrt{n}}$

Lower One-sided $100(1-\alpha)\%$CI:

$\mu \geq \bar{x} - z_\alpha \frac{\sigma ors}{\sqrt{n}}$

CI on μ, σ^2unknown, and n small:

Two-sided $100(1-\alpha)\%$CI:

$\bar{x} - t_{n-1,\alpha/2} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$

CI on $\mu_1 - \mu_2$, σ^2known, or n large:

Two-sided $100(1-\alpha)\%$CI:

$\bar{x} - \bar{y} - z_{\alpha/2} \sqrt{\frac{\sigma_1^2 ors_1^2}{n_1} + \frac{\sigma_2^2 ors_2^2}{n_2}} \leq \mu_1 - \mu_2$

$\leq \bar{x} - \bar{y} + z_{\alpha/2} \sqrt{\frac{\sigma_1^2 ors_1^2}{n_1} + \frac{\sigma_2^2 ors_2^2}{n_2}}$

CI on $\mu_1 - \mu_2$, σ^2unknown, and n small:

Two-sided $100(1-\alpha)\%$CI:

$\bar{x} - \bar{y} - t_{v,\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \leq \mu_1 - \mu_2$

$\leq \bar{x} - \bar{y} + t_{v,\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

where v $= \frac{(w_1 + w_2)^2}{w_1^2/(n_1-1) + w_2^2/(n_2-1)}$

with $w_1 = s_1^2/n_1$ and$w_2 = s_2^2/n_2$

CI on σ^2, $X \sim N(\mu, \sigma^2)$:

Two-sided $100(1-\alpha)\%$CI:

$\chi^2_{n-1,1-\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{n-1,\alpha/2}$

Upper One-sided $100(1-\alpha)\%$CI:

$\sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha}}$

Lower One-sided $100(1-\alpha)\%$CI:

$\sigma^2 \geq \frac{(n-1)S^2}{\chi^2_{n-1,\alpha}}$

Confidence Intervals, Cont.

CI on σ_1^2/σ_2^2, $X \sim N(\mu, \sigma^2)$:

Two-sided $100(1-\alpha)\%$CI:

$\frac{1}{f_{n_1-1,n_2-1,\alpha/2}} \frac{S_1^2}{S_2^2} \leq \frac{\sigma_1^2}{\sigma_2^2}$

$\leq \frac{1}{f_{n_1-1,n_2-1,1-\alpha/2}} \frac{S_1^2}{S_2^2}$

CI on a proportion, p, for n large:

Two-sided $100(1-\alpha)\%$CI:

$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

CI on $p_1 - p_2$, for n_1, n_2 large:

Two-sided $100(1-\alpha)\%$CI:

$\hat{p_1} - \hat{p_2} - z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \leq p_1 - p_2$

$\leq \hat{p_1} - \hat{p_2} + z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

Hypothesis Testing - General

$C =$ Critical Region; $T =$ Test Statistic

Reject H_0 if T is in C.

Type I error: Reject H_0 when it is true

This is a False Positive

Type II error: Fail to reject H_0 when it is false

This is a False Negative

$\alpha =$ Significance Level of the test,

$= \max_{\theta \text{ in } H_0} P_\theta[\text{Type I error}]$

$\beta_\theta = P_\theta[\text{Type II error}]$

$1 - \beta_\theta =$ Power of the test $=$ Power(θ)

A level α test is Uniformly Most Powerful

(UMP) if its power is

$\geq$ the power of any other α test

p-value $=$

$P[T$ at least as extreme as observed $\mid H_0]$

Simple hypothesis:

Determines data distribution, ie. $\mu = \mu_0$

Compound hypothesis:

Determines a set of data distributions, ie.

$\mu \leq \mu_0$

Hypothesis Tests on the Mean

For $X \sim N(\mu, \text{known } \sigma^2)$

Upper One-Sided Test:

reject H_0 if $\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha$

Lower One-Sided Test:

reject H_0 if $\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -z_\alpha$

Two-Sided Test:

reject H_0 if $\left|\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right| > z_{\alpha/2}$

For $X \sim$ Any Dist. and large n

Same as above, but substitute σ with s

For $X \sim N(\mu, \text{unknown } \sigma^2), n$ small

Same as above, but substitute σ with s,

and z_α with $t_{n-1,\alpha \text{ or } \alpha/2}$

Hypothesis Tests on the Variance

For $X \sim N(\mu, \sigma^2)$

Upper One-Sided Test:

reject H_0 if $\frac{(n-1)S^2}{\sigma^2} > \chi^2_{n-1,\alpha}$

Lower One-Sided Test:

reject H_0 if $\frac{(n-1)S^2}{\sigma^2} < \chi^2_{n-1,1-\alpha}$

Two-Sided Test:

reject H_0 if $\frac{(n-1)S^2}{\sigma^2} < \chi^2_{n-1,1-\alpha/2}$

OR $\frac{(n-1)S^2}{\sigma^2} > \chi^2_{n-1,\alpha/2}$

Chi Square Goodness-of-Fit Test

$k =$ num. of groups/categories,

$H_0 =$ a given distribution

$E_i =$ Expected in category i under H_0

$O_i =$ Observed in categroy i

$\sum \frac{(O_i - E_i)^2}{E_i} = \sum \frac{(E_i - O_i)^2}{E_i} = \chi^2_d$

$d = k - 1 - r$

$r =$ num. of parameters estimated from data

Reject H_0 if $\chi^2_d >$ critical value

Contingency Tables

$H_0 =$ data are independent

$H_1 =$ data are not independent

$E_{i,j} =$ expected obs. in row i column j

$O_{i,j} =$ obs. in row i column j

$\sum_{i,j=1}^2 \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} = \chi^2_d$

$d = (r-1)(c-1)$ $r =$ rows $c =$ columns

Reject H_0 if $\chi^2_d >$ critical value

Building the Table:

1. Transform totals into probabilities

   by dividing each cell by n

2. Fill in cells with (totals x probabilities)

3. Mult. each cell by n to get $E_{i,j}$

Neyman Pearson

Most powerful test rejects $H_0 : \theta = \theta_0$

for $H_1 : \theta = \theta_1$ for a given α if

$$\frac{L(\theta_1 | x_1, x_2, ..., x_n)}{L(\theta_0 | x_1, x_2, ..., x_n)} > k$$

To find most powerful test:

1. Determine likelihood functions for H_0, H_1
2. Set up ratio of $L_1/L_0 > k$
3. Simplify, simplify, simplify
4. Isolate x_is as much as possible
5. Get rid of k-side and replace with k*
6. Under H_0, find. k* s.t.

   $P(g(x_i) \geq k^*) = \alpha$

Likelihood Ratio Test

H_0 : data comes from distribution A,

with likelihood L_0

H_1 : data comes from distribution B,

with likelihood L_1, where B generalizes A

$T = 2[ln(L_1) - ln(L_0)] \sim \chi^2_d$

$d =$ num. of extra est. parameters in H_1

Reject H_0 if $T >$ critical value

Information Criteria

$k =$ num. of parameters est. from data

$\hat{L} =$ value of likelihood function at MLE

$n =$ num. of obs. used to find $\hat{L}$

Akaike Information Criteria:

$AIC = 2k - 2ln(\hat{L})$

Bayesian Information Criteria:

$BIC = ln(n)k - 2ln(\hat{L})$

Best model has lowest AIC/BIC