Formula Sheet Applied Statistics

Probability Theory

$E(X+Y)=E(X)+E(Y)$ $\quad E(X-Y)=E(X)-E(Y)$ $\quad E(aX+b)=aE(X)+b$

$var(X)=E(X^2)-(EX)^2$ $\quad var(aX+b)=a^2var(X)$

If X and Y are independent: $\quad var(X+Y)=var(X)+var(Y)$, $\quad var(X-Y)=var(X)+var(Y)$

Distribution	**Probability/Density function**	**Range**	$\boldsymbol{E(X)}$	$\boldsymbol{var(X)}$
Binomial (n,p)	$\binom{n}{x}p^x(1-p)^{n-x}$	$0,1,2,\dots,n$	np	$np(1-p)$
Geometric (p)	$(1-p)^{x-1}p$	$1,2,3,\dots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson (μ)	$\frac{e^{-\mu}\mu^x}{x!}$	$0,1,2,\dots$	μ	μ
Uniform on (a,b)	$\frac{1}{b-a}$	$a\le x\le b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential (λ)	$\lambda e^{-\lambda x}$	$x\ge 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Normal (μ,σ^2)	$\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$\mathbb{R}$	μ	σ^2

Testing procedure in 8 steps

1. Give a probability model of the observed values (the statistical assumptions).
2. State the null hypothesis and the alternative hypothesis, using parameters in the model.
3. Give the proper test statistic.
4. State the distribution of the test statistic if H_0 is true.
5. Compute (give) the observed value of the test statistic.
6. State the test and **a.** Determine the rejection region or
   b. Compute the p-value.
7. State your statistical conclusion: reject or fail to reject H_0 at the given significance level.
8. Draw the conclusion in words.

One Sample

Bounds for Confidence Interval	**Test statistic and distribution under $\boldsymbol{H_0}$**
$\hat{p}\pm c\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, $\quad \Phi(c)=1-\frac{1}{2}\alpha$	Number of successes $X\sim B(n,p_0)$
$\overline{X}\pm c\frac{S}{\sqrt{n}}$, $\quad P(T_{n-1}\ge c)=\frac{1}{2}\alpha$	$T=\frac{\overline{X}-\mu_0}{S/\sqrt{n}}\sim t_{n-1}$
$\left(\frac{(n-1)S^2}{c_2},\frac{(n-1)S^2}{c_1}\right)$, $\quad P(\chi^2_{n-1}\le c_1)=P(\chi^2_{n-1}\ge c_2)=\frac{1}{2}\alpha$	S^2 , where $\frac{(n-1)S^2}{\sigma_0^2}\sim\chi^2_{n-1}$

Prediction interval: $\overline{X}\pm c\sqrt{S^2\left(1+\frac{1}{n}\right)}$

Two Samples

Bounds for Confidence Interval	**Test statistic and distribution under H_0**
$\hat{p}_1 - \hat{p}_2 \pm c\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$, with $\Phi(c) = 1 - \frac{1}{2}\alpha$	$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim N0,1)$, with $\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$
$\overline{X} - \overline{Y} \pm c\sqrt{S^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$, with $S^2 = \frac{n_1 - 1}{n_1 + n_2 - 2}S_X^2 + \frac{n_2 - 1}{n_1 + n_2 - 2}S_Y^2$ and $P\left(T_{n_1+n_2-2} \geq c\right) = \frac{1}{2}\alpha$	$T = \frac{\left(\overline{X} - \overline{Y}\right) - \Delta_0}{\sqrt{S^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$
	$F = \frac{S_X^2}{S_Y^2} \sim F_{n_2-1}^{n_1-1}$

Analysis of categorical variables

* 1 row and k columns: $\chi^2 = \sum_{i=1}^{k} \frac{(N_i - E_0 N_i)^2}{E_0 N_i}$ $\quad (df = k - 1)$
* $r \times c -$ cross table: $\chi^2 = \sum_{j=1}^{c}\sum_{i=1}^{r} \frac{\left(N_{ij} - \hat{E}_0 N_{ij}\right)^2}{\hat{E}_0 N_{ij}}$, with $\hat{E}_0 N_{ij} = \frac{\text{row total} \times \text{column total}}{n}$

  and $df = (r-1)(c-1)$.

Non-parametric tests

* One large (> 40) sample: $Z = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \overset{\text{CLT}}{\sim} N(0,1)$

  Two large samples: $Z = \frac{\left(\overline{X} - \overline{Y}\right) - \Delta_0}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}} \overset{\text{CLT}}{\sim} N(0,1)$
* Sign test: $X \sim B\left(n, \frac{1}{2}\right)$ under H_0
* Wilcoxon`s Rank sum test: $W = \sum_{i=1}^{n_1} R(X_i)$,

  under H_0 with: $E(W) = \frac{1}{2}n_1(N+1)$ and $var(W) = \frac{1}{12}n_1 n_2(N+1)$