STATISTICS HELP CARD

Summary Measures

Sample Mean

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x_i}{n}$$

Sample Standard Deviation

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

Probability Rules

Complement Rule: $P(A^c) = 1 - P(A)$

Addition Rule: $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Conditional Probability: $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$

Events *A* and *B* are independent if $P(A|B) = P(A)$
Events *A* and *B* are independent if $P(A \text{ and } B) = P(A)P(B)$
If *A* and *B* are disjoint events then $P(A \text{ and } B) = 0$

General Discrete Random Variable

Mean $E(X) = \mu = \sum x_i p_i = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k$

Standard Deviation $s.d.(X) = \sigma = \sqrt{\sum (x_i - \mu)^2 p_i}$

Standard Score

$$Standard\ Score = \frac{Observation - Mean}{Standard\ Deviation}$$

Z Score

If *X* follows a Normal distribution
with mean μ and standard deviation σ,
then the random variable $Z = \frac{X-\mu}{\sigma}$
has a $N(0,1)$ distribution

Sample Proportion

$$\hat{p} = \frac{x}{n}$$

Mean $E(\hat{p}) = p$

Standard Deviation $s.d.(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$

Sampling Distribution of $\hat{p}$

If the sample size *n* is large enough
(namely, $np \geq 10$ *and* $n(1-p) \geq 10$),
then the distribution of all possible sample
proportion values is *approximately*

$$N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

Sample Mean

$$\bar{x} = \frac{\sum x_i}{n}$$

Mean $E(\bar{X}) = \mu$

Standard Deviation $s.d.(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

Sampling Distribution of $\bar{X}$

If *X* has Normal distribution
with mean μ and standard deviation σ,
then the distribution of all possible sample
mean values is

$$N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Central Limit Theorem

If *X* follows *any* distribution
with mean μ and standard deviation σ
and the sample size *n* is large enough,
then the distribution of all possible sample
mean values is *approximately*

$$N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

<table>
<tr><th>One Population Proportion</th><th>Difference in Two Population Proportions</th><th>One Population Mean</th><th>Population Mean of Differences</th></tr>
<tr><td>Parameter p
Statistic $\hat{p}$</td><td>Parameter $p_1 - p_2$
Statistic $\hat{p}_1 - \hat{p}_2$</td><td>Parameter μ
Statistic $\bar{x}$</td><td>Parameter μ_d
Statistic $\bar{d}$</td></tr>
<tr><td>Standard Error
$s.e.(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$</td><td>Standard Error
$s.e.(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$</td><td>Standard Error
$s.e.(\bar{x}) = \frac{s}{\sqrt{n}}$</td><td>Standard Error
$s.e.(\bar{x}_d) = \frac{s_d}{\sqrt{n}}$</td></tr>
<tr><td>Confidence Interval
$\hat{p} \pm z^* \times s.e.(\hat{p})$
Conservative Confidence Interval
$\hat{p} \pm \frac{z^*}{2\sqrt{n}}$
Sample Size
$n = \left(\frac{z^*}{2m}\right)^2$
m=desired margin of error</td><td>Confidence Interval
$(\hat{p}_1 - \hat{p}_2) \pm z^* \times s.e.(\hat{p}_1 - \hat{p}_2)$</td><td>Confidence Interval
$\bar{x} \pm t^* \times s.e.(\bar{x})$
$df = n - 1$</td><td>Confidence Interval
$\bar{x}_d \pm t^* \times s.e.(\bar{x}_d)$
$df = n - 1$</td></tr>
<tr><td>Large Sample z-Test
$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$</td><td>Large Sample z-Test
$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$
where $\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$</td><td>One-Sample t-Test
$t = \frac{\bar{x} - \mu_o}{s.e.(\bar{x})}$
$df = n - 1$</td><td>Paired t-Test
$t = \frac{\bar{x}_d - 0}{s.e.(\bar{x}_d)}$
$df = n - 1$</td></tr>
</table>

<table>
<tr><th colspan="2">Difference in Two Population Means</th></tr>
<tr><th>Unpooled (Welch's)</th><th>Pooled</th></tr>
<tr><td>Parameter $\mu_1 - \mu_2$
Statistic $\bar{x}_1 - \bar{x}_2$</td><td>Parameter $\mu_1 - \mu_2$
Statistic $\bar{x}_1 - \bar{x}_2$</td></tr>
<tr><td>Standard Error
$s.e.(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$</td><td>Standard Error
$pooled\ s.e.(\bar{x}_1 - \bar{x}_2) = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
where $s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}$</td></tr>
<tr><td>Confidence Interval
$(\bar{x}_1 - \bar{x}_2) \pm t^* \times s.e.(\bar{x}_1 - \bar{x}_2)$
df from technology **</td><td>Confidence Interval
$(\bar{x}_1 - \bar{x}_2) \pm t^* \times \left(pooled\ s.e.(\bar{x}_1 - \bar{x}_2)\right)$
$df = n_1 + n_2 - 2$</td></tr>
<tr><td>Two-Sample t-Test
$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{s.e.(\bar{x}_1 - \bar{x}_2)} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
df from technology **
**If technology not available, use conservative df = the minimum of $n_1 - 1$ and $n_2 - 1$</td><td>Pooled Two-Sample t-Test
$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{pooled\ s.e.(\bar{x}_1 - \bar{x}_2)} = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
$df = n_1 + n_2 - 2$</td></tr>
</table>

Note: A z-distribution is often used in statistical methods in place of a t-distribution when sample sizes are sufficiently large.

Pearson Correlation and Linear Regression

Pearson Correlation and its square

$$r = \sum \left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$$

$r^2 = \frac{SSReg}{SSTotal}$ where $SSTotal = \sum(y-\bar{y})^2 = SSReg + SSE$

Estimate of σ

$$s = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}}$$

where $SSE = \sum(y-\hat{y})^2 = \sum e^2$

Linear Regression Model

Population Version

Mean: $E(Y|x) = \beta_0 + \beta_1 x$

Individual: $y_i = \beta_0 + \beta_1 x + \varepsilon_i$

where ε_i is $N(0, \sigma)$

Sample Version

Mean: $\hat{y} = b_0 + b_1 x$

Individual: $y_i = b_0 + b_1 x + e_i$

Standard Error of the Sample Slope

$$s.e.(b_1) = \frac{s}{\sqrt{\sum(x-\bar{x})^2}}$$

Confidence Interval for $\boldsymbol{\beta_1}$

$b_1 \pm t^* \times s.e.(b_1)$ *df* = $n-2$

t-Test for $\boldsymbol{\beta_1}$

$t = \frac{b_1 - 0}{s.e.(b_1)}$ *df* = $n-2$

Parameter Estimators

$$b_1 = r\frac{s_y}{s_x}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

Confidence Interval for the Mean Response

$\hat{y} \pm t^* \times s.e.(fit)$ *df* = $n-2$

where $s.e.(fit) = s\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$

Residuals

$$e = y - \hat{y} = observed\ y - predicted\ y$$

Prediction Interval for an Individual Response

$\hat{y} \pm t^* \times s.e.(pred)$ *df* = $n-2$

where $s.e.(pred) = \sqrt{s^2 + \left(s.e.(fit)\right)^2}$

Chi-Square Tests

Test for Goodness of Fit

Expected Count

$$Expected = np_{i0}$$

Test Statistic

$\chi^2 = \sum\frac{(observed-expected)^2}{expected}$ $df = k-1$

Test of Independence

Expected Count

$$Expected = \frac{(row\ total)(column\ total)}{total\ n}$$

Test Statistic

$\chi^2 = \sum\frac{(observed-expected)^2}{expected}$ $df = (r-1)(c-1)$

Properties of a Chi-Square Distribution

A χ^2 random variable has mean = *df* and standard deviation = $\sqrt{2df}$