Summary of Formulae

Sample mean and variance

$$\text{Mean: } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{Variance: } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Probability Rules

$\Pr(A \text{ or } B) = \Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$

$$\Pr(A \text{ and } B) = \Pr(A \cap B) = \Pr(A)\Pr(B|A) = \Pr(B)\Pr(A|B)$$

Random Variables If X and Y are independent random variables, then $W = aX + bY + c$ has:

$$\text{Mean: } \mu_W = a\,\mu_X + b\,\mu_Y + c \qquad \text{Variance: } \sigma_W^2 = a^2\,\sigma_X^2 + b^2\,\sigma_Y^2$$

Discrete Distributions

$$\text{Mean: } \mu_X = \sum_{i=1}^{k} x_i \Pr(X = x_i) \qquad \text{Variance: } \sigma_X^2 = \sum_{i=1}^{k}(x_i - \mu_X)^2 \Pr(X = x_i)$$

Binomial Distribution

$$\mu_X = n\pi \qquad \sigma_X^2 = n\pi(1-\pi) \qquad \Pr(X = x) = \binom{n}{x}\pi^x(1-\pi)^{n-x} \qquad \binom{n}{x} = \frac{n!}{x!\,(n-x)!}$$

If $n\pi \pm 3\sqrt{n\pi(1-\pi)}$ are between 0 and n, then X is approximately normally distributed with mean μ_X and variance σ_X^2.

Normal Distribution A standard normal random variable, Z, has $\mu_Z = 0$ and $\sigma_Z^2 = 1$. To transform a normal random variable X into a standard normal (and vice versa):

$$Z = \frac{X - \mu_X}{\sigma_X} \qquad X = Z\sigma_X + \mu_X$$

Distributions of Statistics

- The mean $\bar{X}$ of a random sample of size n has mean $\mu_{\bar{X}} = \mu_X$ and standard error $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$.
- The sample proportion P computed from a binomial distribution with parameters n and π has a mean of $\mu_P = \pi$ and standard error $\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}}$. If $n\pi \pm 3\sqrt{n\pi(1-\pi)}$ are between 0 and n, then P will be approximately normally distributed.
- The distribution of the difference between two sample means $\bar{X}_1 - \bar{X}_2$ has a mean of $\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$ and a standard error of $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

Contingency Tables

	Factor 2		
Factor 1	Level 1	Level 2	Total
Level 1	a	b	$r_1 = a + b$
Level 2	c	d	$r_2 = c + d$
	$c_1 = a + c$	$c_2 = b + d$	$n = a + b + c + d$

$$\chi^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

- R and C are the number of rows and columns respectively
- $e_{ij} = \frac{r_i c_j}{n}$, where r_i is the ith row total and c_j is the jth column total
- o_{ij} is the observed value in row i column j

Odds ratio: $OR = (a/b)/(c/d) = ad/bc$

Relative risk: $RR = (a/r_1)\,/\,(c/r_2)$

Attributable risk: $AR = a/r_1 - c/r_2$

Confidence Intervals and Hypothesis Tests

All of the $100(1-\alpha)\%$ confidence intervals calculated in this course are of the form:
Estimate $\pm$ multiplier $\times$ standard error
In the table $\bar{x}$, p etc are the values calculated from the samples.

	Estimate	df (ν)	Multiplier	Standard Error
Population mean				
• Random sample, σ_X known	$\bar{x}$	NA	$z_{(1-\alpha/2)}$	$\frac{\sigma_X}{\sqrt{n}}$
• Normal population, σ_X unknown	$\bar{x}$	$n-1$	$t_{(1-\alpha/2,\nu)}$	$\frac{s}{\sqrt{n}}$
• Large random sample ($n \geq 20$), σ_X unknown	$\bar{x}$	$n-1$	$t_{(1-\alpha/2,\nu)}$	$\frac{s}{\sqrt{n}}$
Difference between population means				
• Large random samples (both ≥ 20)	$\bar{x}_1 - \bar{x}_2$	Will be provided	$t_{(1-\alpha/2,\nu)}$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
• Paired difference in random samples from a normal population	$\bar{d}$	$n-1$	$t_{(1-\alpha/2,\nu)}$	$\frac{s_d}{\sqrt{n}}$
Population proportions				
• Population proportion	p	NA	$z_{(1-\alpha/2)}$	$\sqrt{\frac{p(1-p)}{n}}$
• Difference between 2 population proportions	$p_1 - p_2$	NA	$z_{(1-\alpha/2)}$	$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
• Difference between 2 population proportions (hypothesis test)	$p_1 - p_2$	NA	$z_{(1-\alpha/2)}$	$\sqrt{\frac{p^*(1-p^*)}{n_1} + \frac{p^*(1-p^*)}{n_2}}$
	(Use $p^* = \frac{x_1+x_2}{n_1+n_2}$ for hypothesis test)			
Odds ratio, relative risk, attributable risk (see contingency table on previous page for a, b, c and d)				
• Log odds ratio	$\ln(OR)$	NA	$z_{(1-\alpha/2)}$	$\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$
• Log relative risk	$\ln(RR)$	NA	$z_{(1-\alpha/2)}$	$\sqrt{\frac{1}{a} - \frac{1}{r_1} + \frac{1}{c} - \frac{1}{r_2}}$
• Attributable risk – as for the difference between two population proportions with $p_1 = a/r_1$ and $p_2 = c/r_2$				
After ANOVA and Regression				
• Estimate, multiplier and standard error determined from output				

Regression

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ where $\hat{\beta}_1 = \dfrac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.

Standard error of the slope is $s_{\hat{\beta}_1} = \dfrac{s_e}{\sqrt{\sum(x_i - \bar{x})^2}}$ where $s_e = \sqrt{\dfrac{\sum(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\dfrac{\text{RSS}}{n-2}}$

Standard error of a prediction at $X = x_0$ is $PE(\hat{y}_0) = s_e\sqrt{1 + \dfrac{1}{n} + \dfrac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$

ANOVA

$y_{ij} = \mu_i + e_{ij}$ $\qquad$ $\hat{\mu}_i = \bar{y}_{i\cdot}$

$$\underbrace{\text{TSS}}_{\substack{\text{Total sum}\\ \text{of squares}}} = \underbrace{\text{GSS}}_{\substack{\text{Group sum}\\ \text{of squares}}} + \underbrace{\text{RSS}}_{\substack{\text{Residual sum}\\ \text{of squares}}}$$