Selected Formulas

$Range = Max - Min$

$IQR = Q3 - Q1$

Outlier Rule-of-Thumb: $y < Q1 - 1.5 \times IQR$ or $y > Q3 + 1.5 \times IQR$

$$\bar{y} = \frac{\sum y}{n}$$

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

$$z = \frac{y - \mu}{\sigma} \text{ (model based)}$$

$$z = \frac{y - \bar{y}}{s} \text{ (data based)}$$

$$r = \frac{\sum z_x z_y}{n - 1}$$

$$\hat{y} = b_0 + b_1 x \quad \text{where } b_1 = r\frac{s_y}{s_x} \text{ and } b_0 = \bar{y} - b_1\bar{x}$$

$P(\mathbf{A}) = 1 - P(\mathbf{A}^C)$

$P(\mathbf{A} \text{ or } \mathbf{B}) = P(\mathbf{A}) + P(\mathbf{B}) - P(\mathbf{A} \text{ and } \mathbf{B})$

$P(\mathbf{A} \text{ and } \mathbf{B}) = P(\mathbf{A}) \times P(\mathbf{B}|\mathbf{A})$

$$P(\mathbf{B}|\mathbf{A}) = \frac{P(\mathbf{A} \text{ and } \mathbf{B})}{P(\mathbf{A})}$$

A and **B** are independent if $P(\mathbf{B}|\mathbf{A}) = P(\mathbf{B})$. Then $P(\mathbf{A} \text{ and } \mathbf{B}) = P(\mathbf{A}) \times P(\mathbf{B})$

$E(X) = \mu = \sum x P(x)$

$E(X \pm c) = E(X) \pm c$

$E(aX) = aE(X)$

$E(X \pm Y) = E(X) \pm E(Y)$

$Var(X) = \sigma^2 = \sum (x - \mu)^2 P(x)$

$Var(X \pm c) = Var(X)$

$Var(aX) = a^2 Var(X)$

$Var(X \pm Y) = Var(X) + Var(Y)$ if X and Y are independent

Binomial: $P(x) = {}_nC_x p^x q^{n-x}$ $\quad \mu = np \quad \sigma = \sqrt{npq}$

$$\hat{p} = \frac{x}{n} \qquad \mu(\hat{p}) = p \qquad SD(\hat{p}) = \sqrt{\frac{pq}{n}}$$

Sampling distribution of $\bar{y}$:

(CLT) As n grows, the sampling distribution approaches the Normal model with

$$\mu(\bar{y}) = \mu \qquad SD(\bar{y}) = \frac{\sigma}{\sqrt{n}}$$

Inference:

Confidence interval for Parameter = *Estimate* ± *Critical value* × *SE*(*Estimator*)

Test statistic $= \dfrac{Estimate - Parameter}{SE(Estimator)}$ [Replace *SE* by *SD* if latter is known]

Parameter	Estimator	*SD* (Estimator)	*SE* (Estimator)
p	$\hat{p}$	$\sqrt{\frac{pq}{n}}$	$\sqrt{\frac{\hat{p}\hat{q}}{n}}$
$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$	$\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$
μ	$\bar{y}$	$\frac{\sigma}{\sqrt{n}}$	$\frac{s}{\sqrt{n}}$
$\mu_1 - \mu_2$	$\bar{y}_1 - \bar{y}_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
μ_d	$\bar{d}$	$\frac{\sigma_d}{\sqrt{n}}$	$\frac{s_d}{\sqrt{n}}$
σ_ε	$s_e = \sqrt{\frac{\sum(y - \hat{y})^2}{n - 2}}$	(divide by $n - k - 1$ in multiple regression)	
β_1	b_1	(in simple regression)	$\frac{s_e}{s_x\sqrt{n - 1}}$
μ_ν	$\hat{y}_\nu$	(in simple regression)	$\sqrt{SE^2(b_1) \cdot (x_\nu - \bar{x})^2 + \frac{s_e^2}{n}}$
y_ν	$\hat{y}_\nu$	(in simple regression)	$\sqrt{SE^2(b_1) \cdot (x_\nu - \bar{x})^2 + \frac{s_e^2}{n} + s_e^2}$

Pooling: For testing difference between proportions: $\hat{p}_{pooled} = \dfrac{y_1 + y_2}{n_1 + n_2}$

For testing difference between means (when $\sigma_1 = \sigma_2$): $s_p = \sqrt{\dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$

Substitute these pooled estimates in the respective SE formulas for both groups when assumptions and conditions are met.

Chi-square: $\chi^2 = \sum \dfrac{(Obs - Exp)^2}{Exp}$

One-way ANOVA: $SS_T = \sum\sum(\bar{y}_j - \bar{\bar{y}})^2$; $MS_T = SS_T/(k - 1)$

$SS_E = \sum\sum(\bar{y}_{ij} - \bar{y}_j)^2$; $MS_E = SS_E/(N - k)$

$F = MS_T/MS_E$ with $df = (k - 1, N - k)$

- You can make inferences about the difference between two independent means, or about the mean of paired differences using *t*-models.
- You can make inferences about distributions of categorical variables using chi-square models.
- You can make inferences about association between categorical variables using chi-square models.
- You can make inferences about the coefficients in a linear regression model using *t*-models.

Now for some opportunities to review these concepts. Be careful. You have a lot of thinking to do. These review exercises mix questions about proportions, means, chi square, and regression. You have to determine which of our inference procedures is appropriate in each situation. Then you have to check the proper assumptions and conditions. Keeping track of those can be difficult, so first we summarize the many procedures with their corresponding assumptions and conditions on the next page. Look them over carefully . . . then, on to the Exercises!

Assumptions for Inference	And the Conditions that Support or Override them
Proportions (z)	
■ **One sample** 1. Individuals respond independently. 2. Sample is sufficiently large.	1. SRS. 2. Successes $\geq$ 10 and failures $\geq$ 10.
■ **Two sample** 1. Samples are independent of each other. 2. Individual responses in each sample are independent. 3. Both samples are sufficiently large.	1. (Think about how the data were collected.) 2. Both are SRSs OR random allocation. 3. Successes $\geq$ 10 and failures $\geq$ 10 for both samples.
Means (t)	
■ **One sample ($df = n - 1$)** 1. Individuals respond independently. 2. Population has a Normal model.	1. SRS. 2. Histogram is unimodal and symmetric.*
■ **Two independent Samples (df from technology)** 1. Samples are independent of each other. 2. Individual responses in each sample are independent. 3. Both populations are Normal.	1. (Think about the design.) 2. SRSs OR random allocation. 3. Both histograms are unimodal and symmetric.*
■ **Matched pairs ($df = n - 1$)** 1. Each individual is paired with an individual in the other sample; *n* pairs. 2. Individual differences are independent. 3. Population of differences is Normal.	1. (Think about the design.) 2. SRSs OR random allocation. 3. Histogram of differences is unimodal and symmetric.*
Distributions/Association (χ^2)	
■ **Goodness of fit** [df = # of cells −1; one categorical variable, one sample compared with population model] 1. Data are counts of individuals classified into categories. 2. Individuals' responses are independent. 3. Sample is sufficiently large.	1. (Are they?) 2. SRS. 3. All expected counts $\geq$ 5.
■ **Homogeneity** [$df = (r - 1)(c - 1)$; samples from many populations compared on one cateogorical variable] 1. Data are counts of individuals classified into categories. 2. Individuals' responses are independent. 3. Groups are sufficiently large.	1. (Are they?) 2. SRSs OR random allocation. 3. All expected counts $\geq$ 5.
■ **Independence** [$df = (r - 1)(c - 1)$; sample from one population classified on two categorical variables] 1. Data are counts of observations classified into categories. 2. Individuals' responses are independent. 3. Group is sufficiently large.	1. (Are they?) 2. SRSs. 3. All expected counts $\geq$ 5.
Regression with One Predictor and One Response Variable - Both Quantitative (t, $df = n-2$) 1. Form of relationship is linear. 2. Errors are independent. 3. Variability of errors is constant. 4. Errors follow a Normal model.	1. Scatterplot of y against x is straight enough. Scatterplot of residuals against predicted values shows no special structure (e.g. bends). 2. No apparent pattern in plot of residuals against predicted values or against ti[me] (if data collected in time sequence). 3. Plot of residuals against predicted values has constant spread, doesn't "thicken" 4. Histogram of residuals is approximately unimodal and symmetric, or Normal pro[b]ability plot is reasonably straight.*

**Less critical as *n* increases

Note: For all of these procedures, sampling more than 10% of the population compromises independence—but in a good way! Be aware that *P*-values and confide[nce] coefficients calculated in the manner we have discussed become overly conservative (so if your ME or *P*-value is just a bit lacking, get a statistician to help make the appropriate adjustments).

Quick Guide to Inference

Think			Show				Tell?
Inference about?	One sample or two?	Procedure	Sampling Model	Parameter	Estimate	SE	Chapter
Proportions	One sample	1-Proportion z-Interval	z	p	$\hat{p}$	$\sqrt{\frac{\hat{p}\hat{q}}{n}}$	15
		1-Proportion z-Test				$\sqrt{\frac{p_0 q_0}{n}}$	16, 17
	Two independent groups	2-Proportion z-Interval	z	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$	21
		2-Proportion z-Test				$\sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}}, \hat{p} = \frac{y_1 + y_2}{n_1 + n_2}$	21
Means	One sample	t-Interval t-Test	t $df = n - 1$	μ	$\bar{y}$	$\frac{s}{\sqrt{n}}$	18
	Two independent groups	2-Sample t-Test 2-Sample t-Interval	t df from technology	$\mu_1 - \mu_2$	$\bar{y}_1 - \bar{y}_2$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	19
	n Matched pairs	Paired t-Test Paired t-Interval	t $df = n - 1$	μ_d	$\bar{d}$	$\frac{s_d}{\sqrt{n}}$	20
Distributions (one categorical variable)	One sample	Goodness of fit	χ^2 $df = cells - 1$			Test statistic $= \sum \frac{(obs - exp)^2}{exp}$	22
	Many independent samples	Homogeneity χ^2 Test	χ^2 $df = (r - 1)(c - 1)$				
Independence (two categorical variables)	One sample	Independence χ^2 Test					
Association (two quantitative variables)	One sample	Linear Regression t-Test or Confidence Interval for β	t $df = n - 2$	β_1	b_1	$\frac{s_e}{s_x\sqrt{n - 1}}$ (compute with technology)	23
		Confidence Interval for μ_ν		μ_ν	$\hat{y}_\nu$	$\sqrt{SE^2(b_1) \cdot (x_\nu - \bar{x})^2 + \frac{s_e^2}{n}}$	
		Prediction Interval for y_ν		y_ν	$\hat{y}_\nu$	$\sqrt{SE^2(b_1) \cdot (x_\nu - \bar{x})^2 + \frac{s_e^2}{n} + s_e^2}$	
Inference about?	One sample or two?	Procedure	Sampling Model	Parameter	Estimate	SE	Chapter