Unit I Concept Summary

The Who: Cases or observations, "*What you're collecting data on*". Put in rows in StatCrunch. **The What**: The variables, characteristics for each observation. Put in columns.	**Variables** **Quantitative**: Numerical and summary statistics like the mean and median make sense in context. **Categorical**: Classify individuals. Usually words. **Identifier**: A subgroup of categorical, these are labels that cannot really be analyzed statistically. May not be unique.	
Graphs for Quantitative Variables **1. Histogram**, to identify shape, visualize center, spread, unusual features. **2. Boxplot**, shows minimum, quartiles, maximum and any official outliers. **3. Dotplot**, good on StatCrunch to identify certain individuals, like largest, smallest, etc….	**Graphs for Categorical Variables** **1. Pie Charts**, stand alone or side-by-side to look for differences in groups. **2. Bar Charts**, same idea, different display.	**Tables** **1. Frequency Tables**, to count up frequencies or % one variable at a time. **2. Contingency Tables**, to count up frequencies or % two variables at a time.
Describing the Distribution of a Quantitative Variable **1. Shape**, symmetry, skewness, modes **2. Center**, use the mean (roughly symmetric) or median (anything else), depending on shape. **3. Spread**, use the standard deviation (roughly symmetric) or IQR (anything else), depending on shape. **4. Official outliers**, **unusual shapes**, be specific, not generic.	**Analyzing Contingency Tables** **1.** Row totals or column totals over the grand total give **"MARGINAL" proportions**. **2.** Intersections over the grand total give **"AND" proportions**. **3.** Intersections over column totals or row totals give **"CONDITIONAL" proportions**.	**Determining if Two Categorical Variables are Dependent on Each Other** **1.** Using a contingency table, check for differences in marginal distributions and conditional distributions. **2.** Using side-by-side pie charts or stacked bar charts, look for substantial visual evidence of a difference in groups.
Summary Statistics **Sample Mean:** $\bar{y} = \frac{\sum y}{n}$ **Sample Standard Deviation:** $s = \sqrt{\frac{\sum (y-\bar{y})^2}{n-1}}$ **Sample Median:** The value in the middle position. If there isn't one value, take the mean of the two values in the middle positions. Q_1**:** The first quartile, 25th percentile Q_3**:** The third quartile, 75th percentile **IQR** = $Q_3 - Q_1$ P_{60}**:** The 60th percentile, etc…	**Summary Statistics** **Z-Score:** The number of standard deviations a data value lies from the mean. $z = \frac{y-\bar{y}}{s}$ Data values within 2 standard deviations of the mean are **not unusual**. Data values between 2 and 3 standard deviations from the mean **are unusual**. Data values more than 3 standard deviations away from the mean **are rare**.	**Summary Statistics** Official outliers can be determined using the fences: Lower Fence = Q_1 – 1.5(IQR) Upper Fence = Q_3 + 1.5(IQR) **Shifting:** Adding or subtracting the same number to every data value. Measures of position get shifted likewise. Measures of spread remain unchanged. **Scaling:** Multiplying or dividing every data value by the same number. All summary statistics change likewise.

Unit I Concept Summary

Definitions	Sampling Methods
Population: The entire group of interest, usually obtaining data from one is impossible. **Sample:** A subset of the population. In the real world, we typically deal with sample data. **Observational Study:** A data collection method in which no manipulation or attempt is made to affect the outcome. In other words, obtaining data that already exists. **Experiment:** A data collection method in which the experimenter manipulates treatments in order to see the effect on the response variable.	**Simple Random Sample (SRS):** A sample collected in which every individual has an equal chance of being selected and every sample is equally likely to occur. **Stratified Sample:** A sample collected in which the population is first broken down into groups (that are different). Then a SRS is taken from every group. **Cluster Sample:** A sample collected in which the population is first broken down into groups (that are similar). Then clusters are randomly chosen and every member of the cluster is sampled. **Systematic Sample:** A sample collected where every k^{th} individual is selected (like every 10^{th}, e.g.). **Multistage Sample:** A sample collected combining the above methods. **Convenience Sample:** A sample collected with no randomization or usage of the above methods.

Designed Experiments

Response Variable: The variable of interest to the experimenter.

Factors: Variables controlled by the experimenter to see the effects on the response variable. Factor levels are randomly assigned to the experimental units or subjects.

Blocking Factors: Variables controlled by the experimenter to see the effects on the response variable. Blocking factors are not random, but pre-existing in the experimental units or subjects.

Treatment: The combination of factors and levels randomly assigned to the experimental units or subjects.

Statistically Significant: When the results of the experiment are too unusual to attribute to chance, the data / factor / experiment is said to be statistically significant.

Control Group: A treatment group in which no treatment, a baseline treatment, or a placebo treatment is applied.

Replication: Applying a treatment to multiple experimental units or subject.

Lurking Variable: A variable or factor that is the true reason for the results we see in an experiment (not attributed to a specific treatment).

Unit II Correlation and Regression Summary

<u>Conditions for Linear Regression</u>

1. Both variables must be quantitative.
2. The relationship must be linear when viewed on a scatterplot.
3. The spread about the regression line must be equal for all *x*-values.
4. All outliers and influential points must be investigated for accuracy and it must be determined if they are valid data points.

Check a scatterplot and a Residuals vs. *X*-values plot on StatCrunch.

5. Residual plots should have random scatter and equal spread left-to-right.

<u>The Equation of the Linear Regression Line</u>

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

Slope Formula: $b_1 = r\left(\frac{s_y}{s_x}\right)$

y-Intercept Formula: $b_0 = \bar{y} - b_1\bar{x}$

Correlation Formula: $r = \frac{\sum\left(z_x z_y\right)}{n-1}$

<u>If you have data</u>, use StatCrunch features, not formulas.

<u>Interpreting S_e</u>

"On average, our predicted '*y*-variable's values are off by ~ s_e when using *x* = our '*x*-variable'". We want this number to be small.

<u>Interpreting Slope</u>

As the *x* variable increases by one unit, the slope tells us the change in the *y*-variable.

This must be done in the context of the problem.

<u>Interpreting the *y*-Intercept</u>

The *y*-intercept is the value of the *y*-variable when *x* = 0.

Be sure you have data at or very close to *x* = 0, and that the value of the *y*-intercept makes sense in the context of the problem. Otherwise, this value has no interpretable meaning in the context of the problem.

<u>Residuals</u>

1. We want residuals to be small.
2. *e* = Actual *y* – Predicted *y*
3. $e = y - \hat{y}$
4. Residuals can be saved using the StatCrunch checkbox.
5. Be able to interpret residuals in the context of the problem.
6. Data points above the regression line have positive residuals and actual values are greater than the predicted values.
7. Data points below the regression line have negative residuals and actual values are less than the predicted values.
8. Residuals are in the same units as the *y*-variable.

<u>Studentized Residuals</u>

1. Studentized residuals can be saved using the StatCrunch checkbox.

2. Studentized residuals are in "standardized units".

3. Any Studentized residual exceeding ±2 should be flagged. The original row of data should be checked for accuracy and appropriateness.

<u>Cook's Distances</u>

1. Cook's distances can be saved using the StatCrunch checkbox.

2. Any Cook's distance exceeding 4 / *n* should be flagged as an influential point.

3. Any influential points should be checked for accuracy and appropriateness.

4. Do not remove data points from the dataset without good reason. Outliers, data points with large Studentized residuals, or large Cook's distances are investigated.

5. One strategy for valid influential points is to run the regression twice, with and without the points, to see the effect.

<u>R^2</u>

1. $0\% \le R^2 \le 100\%$, with higher being better.
2. R^2 measures the percentage of variation in the *y*-variable accounted for by the linear regression on the *x*-variable.
3. When interpreting, you must put it into the context of the problem.

<u>*r*</u>

1. $-1 \le r \le 1$
2. Correlations from 0 up to 0.5 are weak.
3. Correlations between 0.5 and 0.8 are moderate.
4. Correlations between 0.8 and 1 are strong.

<u>Prediction</u>

1. Plug in the desired *x*-value into the linear equation and solve for the predicted *y*-value.
2. StatCrunch features should be used for the most accuracy.
3. Only predict inside the *x*-range from the data. Predicting outside is extrapolation.

Unit II Probability Summary

Probability Rules	**Uniform Distributions (Continuous)**
1. $P(A^c) = 1 - P(A)$ **2.** P(A or B) = P(A) + P(B) if A and B are disjoint. **3.** P(A and B) = P(A)P(B) if A and B are independent. **4.** P(A|B) = P(A and B) / P(B) **5.** P(A or B) = P(A) + P(B) – P(A and B) for all A, B. **6.** P(A and B) = P(A)P(B|A) for all A, B. **7.** A and B are independent if P(A) = P(A|B). **8.** A and B are disjoint if P(A and B) = 0.	**1.** Between lower and upper boundaries *a* and *b*, all probabilities of intervals of equal length are equally likely. **2.** Rectangular in shape. Height of rectangle is $\frac{1}{b-a}$. **3.** Probability = Area under rectangle. **4.** Percentiles can be found by putting the corresponding area to the left and finding the value on the *x*-axis. **5.** The mean is the midpoint: $\mu = \frac{a+b}{2}$ **6.** The probability function is $f(x) = \frac{1}{b-a}, \quad a \le x \le b$
"At Least" Problems	**Exponential Distributions (Continuous)**
1. In words, determine the complement of what the problem is asking. **2.** Utilize the complement rule. Example given: P(at least one person is Canadian) = 1 – P(no one is Canadian)	**1.** Used to model probabilities for the time between events. **2.** Unimodal and skewed right, short times between events are more likely to occur. **3.** Probability = Area under the curve. **4.** Probabilities for intervals of values and percentiles can be found using Stat → Calculators → Exponential. **5.** Need to be told the mean, μ, which is coincidentally the same as the standard deviation, σ.
(Discrete) Distributions Presented in Table Form	**Normal Distributions (Continuous)**
1. Will be given a list of quantitative outcomes and their respective probabilities. **2.** The mean can be calculated using: $\mu = \sum xP(x)$	**1.** Used to model probabilities with most values piling up near the mean and then tapering off in both directions. **2.** Unimodal and symmetric, defined by the mean, μ, and standard deviation, σ, which will be given in the problem. **3.** Probability = Area under the curve. **4.** Probabilities for intervals of values and percentiles can be found using Stat → Calculators → Normal.

Checking Data For Normality

1. Make a histogram and look for the familiar bell-shaped distribution: unimodal and symmetric.
2. Make a QQ plot and look for the data points to be generally straight on the diagonal.

Binomial Distributions (Discrete)

1. Fixed number of trials, two outcomes per trial, P(Success) = *p* is constant, trials independent.
2. Binomial random variables count up the number of successes, *x*, in *n* trials.
3. $P(X = x) = \left({}_nC_x\right) p^x (1-p)^{n-x}, \quad 0 \le x \le n$
4. $\mu = np$ and $\sigma = \sqrt{np(1-p)}$
5. Can use the StatCrunch calculator rather than the formula, Stat → Calculators → Binomial.

Unit III Inference for Proportions Summary

Conditions for Inference for Proportions	**The Sampling Distribution of p-Hat**
1. The variable or variables are categorical. **2.** The sample(s) are random or at least unbiased. **3.** The sample(s) size(s) are less than 10% of the population(s) size(s). **4.** We have or expect to have at least 10 successes and 10 failures in our sample(s).	As long as the conditions are met, the sample proportion $\hat{p}$ will follow an approximate Normal distribution. The mean and standard deviation of this distribution are as follows: $\mu_{\hat{p}} = p$ $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ This distribution is more for theoretical purposes, because in practice, we are investigating the value of *p*. We won't know it for real world problems, but this is the basis for our confidence intervals and hypothesis tests.

One-Sample Confidence Interval for the Population Proportion

$\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ or use Stat → Proportion Stats → One Sample with data or summary → Confidence Interval.

Margin of Error = $z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \frac{\text{Upper Bound} - \text{Lower Bound}}{2}$

Calculate Sample Size: $n = \frac{z^2 \hat{p}(1-\hat{p})}{(\text{ME})^2}$ with educated guess for $\hat{p}$ or default to 0.50 if necessary.

One-Sample Test for a Population Proportion

Test the hypotheses $\text{H}_0 : p = p_0$ vs $\text{H}_\text{A} : p <$ or $>$ or $\neq p_0$ using the test statistic $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$.

Determine the P-value by shading under the Standard Normal model in the H_A direction.

Or on StatCrunch, Stat → Proportion Stats → One Sample with data or summary → Hypothesis Test.

Reject the null hypothesis in favor of the alternative hypothesis is the P-value is "low" or if the P-value is less than the stated significance level. See page 232 in the notes for the P-value diagram.

Two-Sample Intervals and Tests for a Difference in Proportions

$$(\hat{p}_1 - \hat{p}_2) \pm z\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$\text{H}_0 : p_1 = p_2$ versus $\text{H}_\text{A} : p_1 <$ or $>$ or $\neq p_2$

$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\frac{\hat{p}_{\text{pooled}}(1-\hat{p}_{\text{pooled}})}{n_1} + \frac{\hat{p}_{\text{pooled}}(1-\hat{p}_{\text{pooled}})}{n_2}}}$$

$$\hat{p}_{\text{pooled}} = \frac{x_1 + x_2}{n_1 + n_2}$$

StatCrunch, Stat → Proportion Stats → Two Sample with data or summary → Hypothesis Test or Confidence Interval.

Common Critical *z* Values: 90% → *z* = 1.645 95% → *z* = 1.96 98% → *z* = 2.326 99% → *z* = 2.576

Unit III Inference for Means Summary

<table>
<tr><td>

Conditions for Inference for Means

1. The variable or variables are quantitative and summarized using a mean.

2. The sample(s) are random or at least unbiased.

3. The sample(s) size(s) are less than 10% of the population(s) size(s).

4. We have reason to believe the population is approximately Normal or our sample size(s) are at least size 30.

</td><td>

The Sampling Distribution of *y*-Bar

As long as the conditions are met, the sample mean $\bar{y}$ will follow an approximate Normal distribution. The mean and standard deviation of this distribution are as follows:

$$\mu_{\bar{y}} = \mu_y$$

$$\sigma_{\bar{y}} = \frac{\sigma_y}{\sqrt{n}}$$

This distribution is more for theoretical purposes, because in practice, we are investigating the true value of μ. We won't know μ or σ for real world problems, but this is the basis for our confidence intervals and hypothesis tests.

</td></tr>
</table>

Important

When we run hypothesis tests or compute confidence intervals for means, the true value of σ will be unknown! We substitute the value of *s* from our collected data and as a result, we must use the Student's *t* model instead of the Standard Normal model.

One-Sample Confidence Interval for the Population Mean

$\bar{y} \pm t\left(\frac{s}{\sqrt{n}}\right)$ with df = $n - 1$ or use Stat → T Stats → One Sample with data or summary → Confidence Interval.

Margin of Error = $t\left(\frac{s}{\sqrt{n}}\right) = \frac{\text{Upper Bound} - \text{Lower Bound}}{2}$

Calculate Sample Size: $n = \left(\frac{z(\text{Best Guess for Standard Deviation})}{(\text{ME})}\right)^2$.

One-Sample Test for a Population Mean

Test the hypotheses $\mathrm{H_0}: \mu = \mu_0$ vs $\mathrm{H_A}: \mu <$ or $>$ or $\neq \mu_0$ using the test statistic $t = \frac{\bar{y} - \mu_0}{\left(\frac{s}{\sqrt{n}}\right)}$.

Determine the P-value by shading under the Student's *t* model (df = $n - 1$) in the $\mathrm{H_A}$ direction.

Or on StatCrunch, Stat → T Stats → One Sample with data or summary → Hypothesis Test.

Two-Sample Intervals and Tests for a Difference in Means (Independent Samples, df from technology)

$\mathrm{H_0}: \mu_1 = \mu_2$ versus $\mathrm{H_A}: \mu_1 <$ or $>$ or $\neq \mu_2$

$$t = \frac{(\bar{y}_1 - \bar{y}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \qquad (\bar{y}_1 - \bar{y}_2) \pm t\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

StatCrunch, Stat → T Stats → Two Sample with data or summary → Hypothesis Test or Confidence Interval.

Two-Sample Intervals and Tests for the Mean Difference (Dependent Sample)

Take the differences for each paired data value and run a one-sample *t* test on the column of differences.