Live Classroom Content Review

Final Exam Review/Key Concepts

Statistical Inference

- Statistical methods are used for studying, analyzing, and learning about populations of experimental units (objects about which we collect data from).
- Examples of populations: all people who have diabetes, all orders placed at McDonald's, all females in the state of California, and all truck drivers.
- In real life populations, information about populations is unknown.
  - We used random samples from a population and use the information to make inferences about the population.

Central Limit Theorem

- A powerful result which allows us to use the normal distribution to construct confidence intervals and perform statistical testing relating to the population mean using the sample mean.

- If you have a sample that is sufficiently large, then the distribution of the sample mean will be approximately normally distributed with
  - a mean of μ (equal to the population mean)
  - standard deviation of $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ (equal to the population standard deviation divided by the square root of the sample size)

- As the sample size increase, the variability decreases.

Central Limit Theorem

Properties of the Normal Distribution

- Symmetric
- Area sums to 1

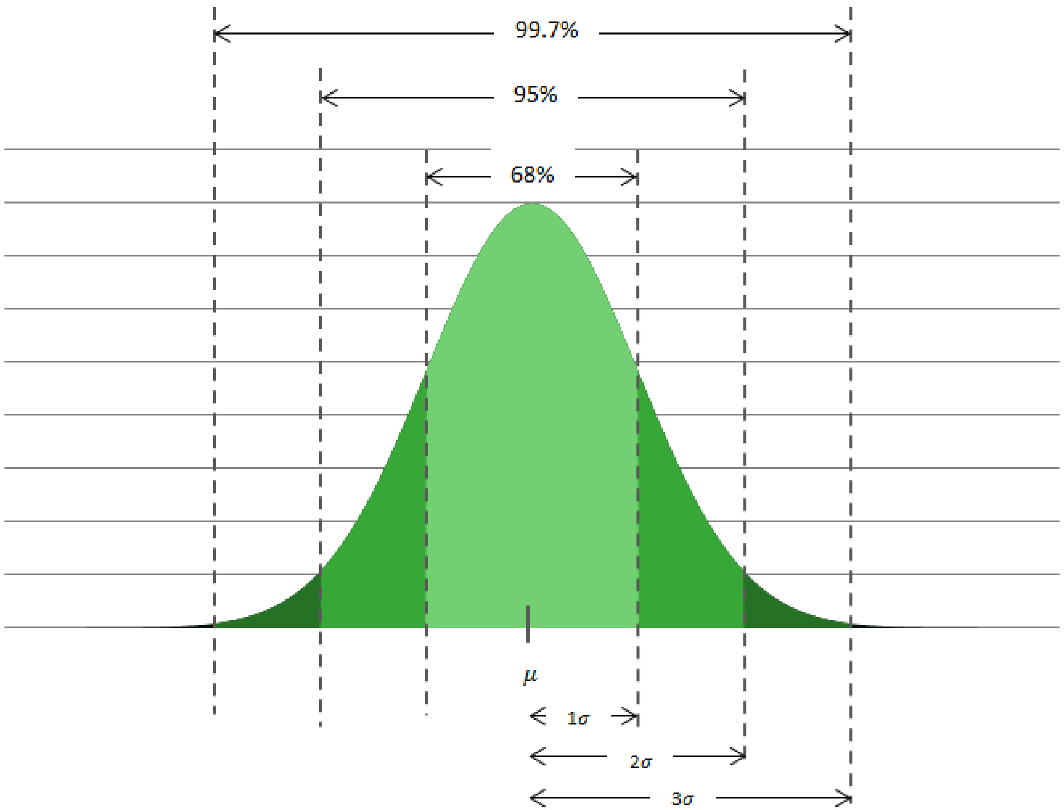


Table A. Standard Normal Probabilities (continued)

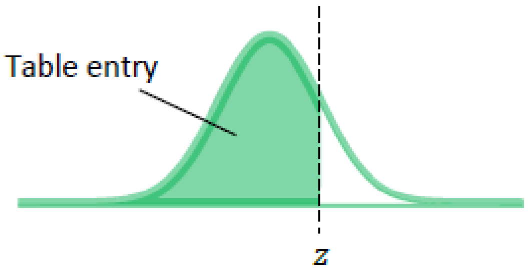


Table entry for z is the area under the standard Normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.00	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.10	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.20	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.30	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.40	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.50	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.60	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.70	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.80	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.90	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.00	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.10	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.20	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.30	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.40	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.50	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.60	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.70	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.80	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.90	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767

Different Tests we have Learned

- When to use which analysis
  - One-sample test for means
  - Two-sample test for means
  - Correlation/Simple Linear Regression
  - Multiple Linear Regression
  - ANOVA
  - ANCOVA
  - One-sample test for proportions
  - Two-sample test for proportions
  - Logistic Regression

Tests for Means

- **One-sample test for means:** Testing whether the population mean is equal to a specific value.
  - Z or T test depending on whether population sd is known and what the sample size is.
  - Use t-test when you don't know the population standard deviation AND when your sample size is less than 30.

- **Two-sample test for means:** Testing whether the mean of one population is equal to the mean in another population.
  - We used a T test.

- **Correlation/ Simple Linear Regression:** Testing for a linear association between two quantitative variables.
  - T or F test.

Tests for Means

- **Multiple Linear Regression:** Testing whether there is a linear relationship between a continuous dependent variable and a set of explanatory variables.
  - **(1) Global test – F test**
  - **(2) If global test significant – test each explanatory variable controlling for other variables in the model. We used T-tests for these.**

Tests for Means

- **One-way ANOVA:** Testing whether the means differ across groups (one variable).
  - Global F-test.
  - Pairwise comparisons between groups with T-tests.
  - Relationship with Regression – dummy variables
  - ANCOVA: adjusting for a continuous or categorical variable.

- **Two-way ANOVA:** Testing whether the means differ across groups (two variables/factors).
  - Test of interaction, global test, test of main effects.
  - If interaction is significant → stratify
  - ANCOVA: adjusting for a continuous or categorical variable.

5 Step Recipe for Testing

- Set up the hypotheses and select the α level.
- Select the appropriate test statistic.
- State the decision rule.
  - Critical value approach.
  - P-value approach.
- Compute the test statistic and associated p-value.
- State your conclusion.

Be able to perform any of these steps on the final exam! You will be asked to interpret R output from some of the tests we have learned about.

Other Key Concepts

- Interpretation of confidence intervals
  - We are 95% confident that the underlying…
  - Relationship with testing (across different types of analyses)
    - A level α significance test rejects the null hypothesis H_0: $\mu = \mu_0$ when the value of μ_0 is *not included* in the 1- α confidence interval for μ.
    - A level α significance test fails to reject the null hypothesis H_0: $\mu = \mu_0$ when the value of μ_0 is *included* in the 1- α confidence interval for μ.
    - ** The conclusion of a **two-sided** significance test (whether or not the null hypothesis is rejected) at the α level of significance can be determined by checking if the "null" value as specified by the null hypothesis is contained within the **1-** $\boldsymbol{\alpha}$ confidence interval. **

Other Key Concepts

- Multiple comparison procedures (such as in the ANOVA setting.)
  - When we do lots of comparisons, we increase our risk of a Type I error
  - To avoid this, we need to make it harder to reject. This can be accomplished numerous ways (all equivalent):
    - Increase the p-value.
    - Decrease the significance level used for each test.
    - Increase the critical value used for decision rule.
  - Various methods for doing this, most simple is Bonferroni.
    - Divide the significance level by the number of tests.
    - OR multiply each p-value by the number of tests.

Other Key Concepts

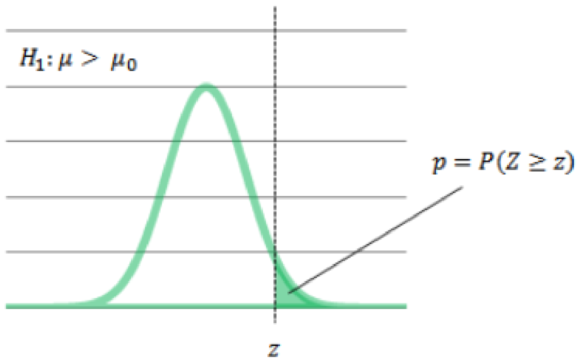
- Interactions
  - Types
    - **Quantitative** – direction of effects are consistent, but the magnitude of the effects of one factor across the other are different
    - **Qualitative** – the direction of the effects are opposite rather than just varying in magnitude
  - What to do about them
    - If the effect of one variable on another depends on the level of a third variable, then you can't just "adjust" for this.
    - Requires stratifying by the level of the third variable so that the association is accurate in each case.

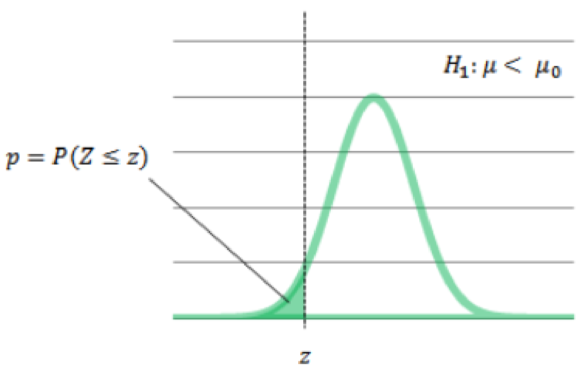
Other Key Concepts

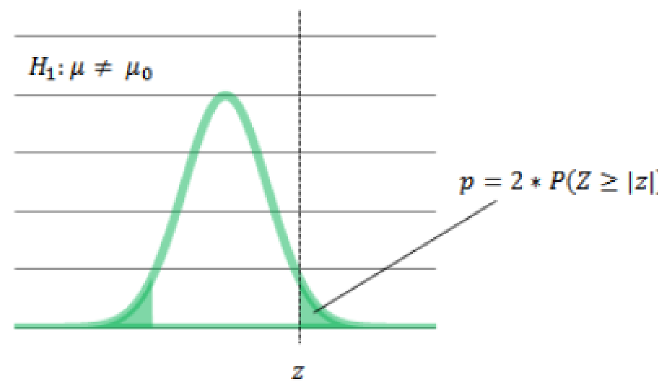
- Interpreting output from regression (logistic or linear)
  - How to identify which values go with which tests
    - Which values are for the global tests and which for individual assessment of each predictor
  - What the underlying null and alternative hypotheses are for both the global test and the individual tests
  - How to interpret the values from the output (beta estimates, R-squared, p-values)

Key Calculations

- **Calculating p-values**
  - One-sided vs Two-sided
  - Using tables, based on z, t, or F distribution (but remember F distribution is not symmetric!)







Key Calculations

- ANOVA table dependencies
  - How df is calculated for each type of analysis (regression, ANOVA)
  - Be able to fill in ANOVA table when some pieces are missing.

	SS (Sum of Squares)	df (degrees of freedom)	MS (Mean Square)	F-statistic	p-value
Regression	Reg SS	Reg df	Reg MS = Reg SS/Reg df	F=Reg MS/Res MS	P($F_{\text{Reg df,Res df, }\alpha}$>$F$)
Residual	Res SS	Res df	Res MS = Res SS/Res df		
Total	Total SS = Reg SS + Res SS				

Key Calculations

- Z-score
  - $z = \frac{x-\mu}{\sigma/\sqrt{n}}$
- Prediction using regression equation
  - Linear Regression

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

  - Logistic Regression

$$\hat{p} = \frac{e^{\widehat{\beta}_0+\widehat{\beta}_1 x_1+\widehat{\beta}_2 x_2+\cdots+\widehat{\beta}_k x_k}}{1 + e^{\widehat{\beta}_0+\widehat{\beta}_1 x_1+\widehat{\beta}_2 x_2+\cdots+\widehat{\beta}_k x_k}}$$

- Residuals

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

Key Calculations

- Odds ratio for 1-unit or x-unit increase in explanatory variable in Logistic Regression:

$$\widehat{OR}_{x_a\ versus\ x_b} = e^{\hat{\beta}_1 x_a - \hat{\beta}_1 x_b} = e^{\hat{\beta}_1 (x_a - x_b)}$$

- Odds Ratio Confidence Interval:

$$\widehat{OR}_{x_a\ versus\ x_b} = e^{\left(\hat{\beta}_1 \pm z_{\frac{\alpha}{2}} * SE_{\hat{\beta}_1}\right)(x_a - x_b)}$$

Key Notes to Remember in Linear Regression

- Relationship between beta estimate and correlation coefficient in Simple Linear Regression
  - Positive association ($r > 0, \beta > 0$)
    - As one variable goes up, the other goes up
    - As one variable goes down, the other goes down
  - Negative association ($r < 0, \beta < 0$)
    - As one variable goes up, the other goes down
    - As one variable goes down, the other goes up

Key Notes to Remember in Linear Regression

- Interpretation of beta estimates (in SLR vs MLR)
- Purpose of diagnostic plots
  - **Residual Plots:** Assess regression assumptions (linearity, constant variance) and to identify outliers
  - **Histogram of the Residuals:** Assess regression assumptions (normality) and to identify observations with large residuals