2019 Summer STAT 22000 Final Exam Study Guide

The final exam is NOT **cumulative**.

Binomial distributions (section 3.4, Slides L07binomial.pdf) For $X \sim Bin(n, p)$,

- $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$, for $k = 0, 1, \ldots, n$
- $\mathrm{E}(X) = np$, and $\mathrm{SD}(X) = \sqrt{np(1-p)}$
- If $np \geq 10$ and $n(1-p) \geq 10$, then $Bin(n, p) \approx N\left(\mu = np,\ \sigma = \sqrt{np(1-p)}\right)$
- Assumptions:
  - Only two possible outcomes in each trial (Please first specify what constitutes a trial in the context of the problem.)
  - The number of trials n must be fixed in advance
  - The probability that the event occurs, p, must be the same from trial to trial
  - The trials must be independent
- Continuity correction

CLT and Sampling distributions (section 4.1&4.4, Slides L08CLT.pdf)

- standard error for the sample mean
- what is the sampling distribution of a sample mean?
- when can one use the CLT?
- sample problems: Exercise 4.33, 4.35, 4.37, 4.39, 4.41, Problem 2-3 in HW8 and Problem 1 in HW9

Overview of Confidence Intervals (section 4.2, Slides L09ConfIntvl.pdf)

- interpretation of confidence intervals
- what's the thing that has a 95% probability to happen?
- conditions required for using a confidence interval on p.178
- marginal of error = half of the width of CI
- confidence level
- sample problems: Exercise 4.13, Problem 2-3 in HW9

Overview of Hypothesis Testing (section 4.3, Slides L10HypTests.pdf, L11.pdf)

- H_0, H_a, test statistic, p-value
- framework of hypothesis testing: assuming H_0 is true, then evaluate the test results to determine if there is enough evidence to reject H_0
- interpretation of p-value: $\mathrm{P}(\text{data} \mid \mathrm{H}_0 \text{ is true})$, not $\mathrm{P}(\mathrm{H}_0 \text{ is true} \mid \text{data})$
- Type 1 error = falsely reject a true H_0, Type 2 error = failing to reject a false H_0
- critical value approach and p-value approach
- significance level = chance of making a Type 1 error
- failing to reject H_0 doesn't prove H_0 to be true
- H_0 and H_a are always statements about population(s), not about samples, eg, Exercise 4.19
- relationship between hypothesis testing and confidence intervals
- statistical significance doesn't mean practical importance
- Don't take the 0.05 significance level too seriously. A p-value of 0.049 or 0.051 do not differ much in the strength of evidence against H_0
- hypothesis testing cannot tell us if data were collected properly or if the design of a study was flawed
- sample problems: Exercise 4.19, 4.21, 4.24, 4.29, 4.31, 4.32(a)(b)(c)

One-sample, Two-sample, paired data problems about population means (Section 5.1-5.3, Slides L12OneSampleMean.pdf, L13TwoSampleMeans.pdf, L14Paired.pdf)

- t-distributions
- one sample t-test, t-interval (check for skewness and outliers before using t-procedures)
  H$_0$: $\mu = \mu_0$: test statistic $t = \dfrac{\bar{x} - \mu_0}{\text{SE}}$ where SE $= s/\sqrt{n}$
  CI for μ: $\bar{x} \pm t^*$ SE $df = n - 1$, $t^* =$ `qt(alpha/2, df, lower.tail=F)`
- comparison: t v.s. normal, t-intervals v.s z-intervals, t-tests v.s. z-tests
- two sample t-tests, t-intervals (always assume unequal population SDs, and check for skewness and outliers)
  H$_0$: $\mu_1 = \mu_2$: test statistic $t = \dfrac{\bar{x}_1 - \bar{x}_2}{SE}$, CI for $\mu_1 - \mu_2$: $(\bar{x}_1 - \bar{x}_2) \pm t^*$ SE,
  where SE $= \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, df $= \min(n_1 - 1, n_2 - 1)$, $t^* =$`qt(alpha/2, df, lower.tail=F)`
- analysis of paired data = one-sample problem on the differences of paired observations.
  H$_0$: $\mu_1 = \mu_2$: test statistic $t = \bar{d}/\text{SE}$, CI for $\mu_1 - \mu_2$: $\bar{d} \pm t^*$ SE
  where $d_i = x_{1i} - x_{2i}$, SE $= s_d/\sqrt{n}$, df = # of pairs -1, $t^* =$`qt(alpha/2, df, lower.tail=F)`
  When checking conditions, just check whether the differences are skewed or having any outlier(s).
- when to use a one-sample, paired, or two-sample analysis
- sample problems: Exercise 5.11, 5.15, 5.17, 5.25, 5.33 on p.259-266 of the textbook (Brief answers can be found on p.416-417 of the textbook.)

One- and two-sample problems about proportions (Section 6.1-6.2, Slides L15Proportions.pdf)

- one sample $100(1 - \alpha)\%$ C.I. for a single proportion (condition: $n\widehat{p}$ and $n(1 - \widehat{p})$ both ≥ 10)
  $$\widehat{p} \pm z^* \times \sqrt{\widehat{p}(1 - \widehat{p})/n}, \quad \text{where } z^* = \texttt{qnorm(alpha/2, lower.tail=F)}$$
- sample size required to control the margin of error of a $100(1 - \alpha)\%$ CI at m:
  $$n \geq \left(\frac{z^*}{m}\right)^2 \widehat{p}(1 - \widehat{p})$$
- one sample test for a single proportion
  test statistic for H$_0$: $p = p_0$ is $z = \dfrac{\widehat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$.
  condition: np_0 and $n(1 - p_0)$ both ≥ 10
- C.I. for the difference of two proportions $p_1 - p_2$:
  $$(\widehat{p}_1 - \widehat{p}_2) \pm z^* \times \sqrt{\frac{\widehat{p}_1(1 - \widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1 - \widehat{p}_2)}{n_2}}, \quad \text{where } z^* = \texttt{qnorm(alpha/2, lower.tail=F)}$$
  condition: $n_1\widehat{p}_1$, $n_2\widehat{p}_2$, $n_2(1 - \widehat{p}_2)$ and $n_2(1 - \widehat{p}_2)$ all ≥ 10
- test for the difference of two proportions (note we use the pooled SE here)
  test statistic for H$_0$: $p_1 = p_2$ is
  $$z = \frac{(\widehat{p}_1 - \widehat{p}_2) - 0}{\sqrt{\widehat{p}(1 - \widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \text{ where } \widehat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$
  condition: $n_1\widehat{p}$, $n_2\widehat{p}$, $n_2(1 - \widehat{p})$ and $n_2(1 - \widehat{p})$ all ≥ 10
- sample problems: Exercise 6.1, 6.5, 6.9, 6.15, 6.23, 6.25, 6.27, 6.29 on p.312-319 of the textbook (Brief answers can be found on p.419-421 of the textbook.)

Correlation (Section 7.1.4, Slides L16Correlation.pdf)

- correlation reflects the direction and strength of linear association
- visual estimation of correlation from a scatterplot
- when will $r = 1$ or -1
- r doesn't change if interchanging x & y or if the unit of x or y is changed
- limitation of r: very sensitive to outlier, may not sensible if the data is clustered, cannot reflect strength of non-linear association
- correlation is not causation

Regression (Section 7.1-7.4, Slides L17Regression.pdf, L18SLRModels.pdf)

- the idea of least square method
- least square regression line:
  slope $= r \times$ (SD of y)/(SD of x),
  intercept = (mean of y)$-$ (slope) $\times$ (mean of x)
- interpretation of the slope and the intercept of the least square regression line
- residual = observed $y-$ predicted y = the signed vertical distance (not the shortest distance) from the data point to model line.
- Extrapolation (prediction beyond the range of the data) is usually unreliable
- For LS regression, residuals add up to zero, and have 0 correlation with the explanatory variable.
- R^2 = R-squared $= r^2$ = proportion of variation in the response y that can be explained by the explanatory variable
- Regression treats x and y differently.
  The LS regression line that predicts y from x and the one that predicts x from y are different.
  The LS regression line that predicts y from x can only predict y from x, not x from y.
- assumption of simple linear regression model: (independence, linearity, constant variability, normality)
- checking model assumptions using residual plots and histograms of residuals
- SE for the intercept and slope can be obtained from R summary output
- estimate for sigma
- using R summary output to find C.I.s and perform t-tests for the intercept and slope
- outlier, influential point, high leverage point
- sample problems: Exercise 7.1, 7.7, 7.13, 7.19, 7.21, 7.25, 7.27, 7.31, 7.37, 7.41, and the exercises posted on Canvas.