EXAM 2 REVIEW

Definition. A **random variable** X assigns a number to each outcome in the sample space S.

(1) All random variables have a **cumulative distribution function (CDF)**: $F(x) = P(X \leq x)$.
(2) A discrete random variable has a **probability mass function (PMF)**: $m(x) = P(X = x)$.
(3) A continuous random variable has a **probability density function (PDF)** $f(x)$ such that for any numbers a and b (with $a \leq b$)

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

Definition. Expected value (or **mean**):

(1) If X is a discrete RV with PMF $m(x)$, then $\mu = E(X) = \sum_x xm(x)$
For any function g, $E(g(X)) = \sum_x g(x)m(x)$
(2) If X is a continuous RV with PDF $f(x)$, then $\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$
For any continuous function g, $E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx$

Definition. The **variance** of X: $\sigma^2 = \text{Var}(X) = E\left[(X-\mu)^2\right] = E(X^2) - [E(X)]^2$. The **standard deviation** of X: $\sigma = \sqrt{\sigma^2}$.

Theorem. *For any random variable X and any constants a and b:*
(1) $E(aX + b) = aE(X) + b$ and
(2) $Var(aX + b) = a^2\, Var(X)$.

Theorem. $E(X_1+X_2+\cdots+X_n) = E(X_1)+E(X_2)+\cdots+E(X_n)$ *for any random variables $X_1, X_2, \ldots X_n$. If $X_1, X_2, \ldots X_n$ are independent, then it also follows that $Var(X_1+X_2+\cdots+X_n) = Var(X_1)+Var(X_2)+\cdots+Var(X_n)$.*

Definition. A **random sample of size** n is a set of independent identically distributed (iid) random variables $X_1, X_2, \ldots X_n$. Some **sample statistics**:

(1) The **sample total** $T = \sum_{i=1}^{n} X_i$
(2) The **sample mean**: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
(3) The **sample variance**: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$

Theorem. *For any random sample from a population with mean μ and variance σ^2:*
(1) $E(T) = n\mu$ and $Var(T) = n\sigma^2$
(2) $E(\overline{X}) = \mu$ and $Var(\overline{X}) = \frac{\sigma^2}{n}$
(3) $E(S^2) = \sigma^2$

Definition. The **standard error** of the sample mean is $s/\sqrt{n}$

Definition. A sample statistic $\hat{X}$ is an **unbiased estimator** of population parameter ρ if $E(\hat{X}) = \rho$.

Theorem. *If $\overline{X}$ is a the mean of a random sample from a normally distribution population, then $\overline{X}$ is normally distributed (with mean and variance given in the last theorem).*

Theorem (Central Limit Theorem)**.** *If $\overline{X}$ is a the mean of a large random sample from any population, then $\overline{X}$ is approximately normally distributed (with mean and variance given in the theorem above).*

Date: April 2, 2020.

A. Some confidence (and prediction) intervals

1. $100(1-\alpha)\%$ CI for μ (known σ, normal population or large sample): $\overline{x} \pm z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$

2. $100(1-\alpha)\%$ CI for μ (normal population or large sample): $\overline{x} \pm t_{\alpha/2,n-1}\dfrac{s}{\sqrt{n}}$

3. $100(1-\alpha)\%$ **prediction interval** for μ (normal population): $\overline{x} \pm t_{\alpha/2,n-1}\sqrt{\dfrac{s^2(n+1)}{n}}$

4. $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$ (normal populations with the same variance):

$$\overline{x}_1 - \overline{x}_2 \pm t_{\alpha/2,n_1+n_2-2}\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

where $s_p^2 = \dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$ is the pooled estimator of the common variance

5. $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$ (normal populations with difference variances):

$$\overline{x}_1 - \overline{x}_2 \pm t_{\alpha/2,\nu}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

where $\nu \approx \dfrac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}$ (round down if you need an integer)

6. $100(1-\alpha)\%$ CI for a proportion θ (large sample; $n\overline{x}$ and $n(1-\overline{x})$ are both at least 10):

$$\overline{x} \pm z_{\alpha/2}\sqrt{\frac{\overline{x}(1-\overline{x})}{n}}$$

B. Some hypothesis tests

Definition. When conducting a hypothesis test there are two types of error:

(1) **Type I error** is rejecting H_0 when H_0 is actually true.
(2) **Type II error** is failing to reject H_0 when H_0 is actually false.

The **significance level** of a test is the probability of a type I error.

Method. For tests about a proportion, the null hypothesis should be $\boxed{H_0 : \theta = \theta_0}$ and our test statistic is the sample total T. Under H_0, $T \sim \text{binom}(n, \theta_0)$.

Method. For tests about the mean of a population (with unknown variance) where either the sample is large or the population is normally distributed, the null hypothesis should be $\boxed{H_0 : \mu = \mu_0}$ and our test statistic is $\boxed{T = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}}$, which has a t-distribution with $n - 1$ degrees of freedom.

Method. For tests about the difference between the means of two populations (with unknown variances), the null hypothesis should be $\boxed{H_0 : \mu_1 - \mu_2 = \delta_0}$ and our test statistic is one of the following. We have independent random samples of sizes n_1 and n_2 with means $\overline{x}_1$ and $\overline{x}_2$ and sample variances s_1^2 and s_2^2.

A) $\boxed{t = \dfrac{\overline{x}_1 - \overline{x}_2 - \delta_0}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}}$

Uses a pooled estimator for variance: $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$

Has a t-distribution with $n_1 + n_2 - 2$ degrees of freedom

Appropriate when **populations have the same variance**, all sample sizes if the populations are normal, otherwise large samples only

B) $\boxed{t = \dfrac{\overline{x}_1 - \overline{x}_2 - \delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}}$

Has a t-distribution with ν degrees of freedom where $\nu = \dfrac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$

Appropriate when **populations have different variances**, all sample sizes if the populations are normal, otherwise large samples only

Method. Hypothesis tests involving variances.

A) $\boxed{f = \dfrac{s_1^2}{s_2^2}}$ for tests of $\boxed{H_0 : \sigma_1^2 = \sigma_2^2}$

Has an F-distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom (order matters)

Appropriate only when populations are normally distributed (or samples are very large)

R CDF: `pf(`f`,` $n_1 - 1$`,` $n_2 - 1$`)`

B) $\boxed{x = \dfrac{(n-1)s^2}{\sigma_0^2}}$ for tests of $\boxed{H_0 : \sigma^2 = \sigma_0^2}$

Has a chi-square distribution with $n - 1$ degrees of freedom

Appropriate when populations are normally distributed or the sample is large

R CDF: `pchisq(`x`,` $n - 1$`)`

C. Practice problems

1. I have collected the heights of 71 adult humans, more or less at random. This sample has mean $\overline{x} = 68.28$ and standard deviation of $s = 3.60$. Calculate a 90% confidence interval for the mean height of an adult human.

2. A January, 2016 YouGov poll of 388 likely South Carolina Democratic primary voters found that 60% would vote for Clinton. Calculate a 95% confidence interval for the true percentage of Clinton voters in the South Carolina Democratic primary (which was held on Saturday, February 27 and in which Clinton got 73.5% of the votes).

3. The article "Orchard Floor Management Utilizing Soil-Applied Coal Dust for Frost Protection" reports the following values for soil heat flux of $n = 8$ plots covered with coal dust:

34.7 35.4 34.7 37.7 32.5 28.0 18.4 24.9

The mean soil heat flux for plots covered only with grass is 29.0. Assuming that soil heat flux is normally distributed, does this data evidence that coal dust is effective in increasing mean soil heat flux?

4. It is thought that roughly $\frac{2}{3}$ of all people have a dominant right foot and $\frac{2}{3}$ have a dominant right eye. Do people also kiss to the right? The article "Human Behavior: Adult Persistence of Head-Turning Asymmetry" reported that in a random sample of 124 kissing couples, 80 of the couples tended to lean more to the right than left. Does this result suggest that more than half of all couples lean right when kissing? Does this result provide evidence against the hypothesis that $\frac{2}{3}$ of all kissing couples lean right?

5. Suppose that a population is uniformly distributed on the interval $[1, \beta]$. Let $X_1, X_2, \ldots, X_{16}$ be a random sample from this population. Find an unbiased estimator for β.

Challenge. Sometimes people conducting surveys want to ask questions with potentially embarrassing answers. They may use something called a *randomized response technique* to make sure that survey participants have no reason to lie. For example, suppose you wanted to determine how widespread cheating is on campus. You could print the phrase "I have cheated on a test" on 4 cards and "I have never cheated on a test" on 6 cards. Then you could ask each survey participant to select a random card and tell you if the statement on the card is true or false *without showing you the card*.

a) Suppose 55% of those surveyed said the statement on the card was true. What is your estimate of the percentage of Gonzaga students who have cheated?
b) Suppose 60% of those surveyed said the statement on the card was true. What is your estimate of the percentage of Gonzaga students who have cheated?
c) Suppose 65% of those surveyed said the statement on the card was true. What is your conclusion?