I. Results related to normal distribution

1. Expected value and variance.
   (a) $E(aX+bY) = aEx + bEY$, $\text{Var}(aX+bY) = a^2 \text{Var}X + b^2 \text{Var}Y$ provided $X$ and $Y$ are independent.

2. Normal distributions:
   (a) $Z \sim N(0, 1)$
   (b) $X \sim N(\mu, \sigma^2)$, then $\frac{X-\mu}{\sigma} \sim N(0, 1)$
   (c) if $X \sim N(\mu_x, \sigma_x^2)$, $Y \sim N(\mu_y, \sigma_y^2)$, and they are independent, then
      
      
      
      
      
      
      
      
      
      In particular, if $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$, $i=1,2, \ldots, n$, 
      
      
      
      
      

3. The central limit theorem: If $X_i \overset{iid}{\sim} f(x)$, $i = 1, \ldots, n$, $\mu = EX$, $0 < \sigma^2 = \text{Var}X < \infty$, then when $n$ is large, $\sum_{i=1}^{n} X_i \approx \sim N(n\mu, n\sigma^2)$ or $\overline{X} \approx \sim N(\mu, \sigma^2/n)$.

4. Binomial Distribution (and sample):
   (a) Proportion of “successes” in $n$ independent (Bernoulli) trials, each with probability $p$ of “success”. Has probability function
      
      
      
      (population) mean = $np$, population $SD = \sqrt{(npq)}$
   (b) Sample estimate of $p$ is $\hat{p} = k/n$. SD of this estimate is $\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$. This is estimated by
       $\hat{\sigma}_{\hat{p}} = \sqrt{\frac{pq}{n}} = \text{Standard Error}$
   (c) When $n$ is large $\hat{p}$ is approximately $N(p, \sigma_{\hat{p}})$

4. $\chi^2_n$ distribution.
   (a) Let $Z_i \overset{iid}{\sim} N(0, 1)$, $i = 1, \ldots, n$, then
      
      
      (b) If $U \sim \chi^2_n$, then
       i. $U$ has Gamma($\frac{n}{2}, \frac{1}{2}$). $EU=n$, $\text{Var}U=2n$.
       ii. $X \sim \text{Gamma}(\alpha, \lambda)$, $f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$, $x>0$.
   (c) If $U_i \overset{iid}{\sim} N(\mu_i, \sigma_i^2)$, then $\sum_{i=1}^{n} \left( \frac{U_i - \mu_i}{\sigma_i} \right)^2 \sim \chi^2_n$. 

2-1
5. \( t_n \) distribution
   
   (a) If \( Z \sim N(0, 1) \), independent of \( U \sim \chi^2_n \), then
   \[
   \frac{Z}{\sqrt{U/n}} \sim t_n
   \]
   
   (b) The density function of \( t_n \) is symmetric about 0.
   
   (c) When the degrees of freedom \( n \to \infty \), then \( f_{t_n}(t) \Rightarrow \text{Standard Normal} \)

6. F_{m,n} distribution. If \( U \sim \chi^2_m \) independent of \( V \sim \chi^2_n \), then
   \[
   \frac{U}{m} \sim F_{m,n}.
   \]

7. The sample mean and the sample variance.
   Suppose \( X_i \overset{iid}{\sim} f \) with mean \( \mu \) and variance \( \sigma^2 \).
   
   (a) Sample mean \( \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \)
   
   Sample variance \( s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1} = \frac{\sum X_i^2 - n(\bar{X})^2}{n-1} \).
   
   \( E(\bar{X}) = \mu, \ Var(\bar{X}) = \sigma^2/n, \ E(s^2) = \sigma^2 \)
   
   (b) If \( X_i \overset{iid}{\sim} N(\mu, \sigma^2) \), then
      
      i) \( \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1} \)
      
      ii) \( \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1} \)

II. Confidence Intervals and Hypothesis testing.

Let \( X_1, X_2, \ldots, X_n \) be a random sample from \( f(x, \theta) \).

1. \( (T_1(X_1, X_2, \ldots, X_n) < \theta < T_2(X_1, X_2, \ldots, X_n)) \) is a 100(1 - \( \alpha \))% CI of \( \theta \) if
   \[
   P(T_1(X_1, X_2, \ldots, X_n) < \theta < T_2(X_1, X_2, \ldots, X_n)) = 1 - \alpha.
   \]

2. Hypotheses:
   - The null hypothesis, \( H_0 \), statements we are against.
   - The alternative hypothesis, \( H_a \), or \( (H_1, H_A, H_\alpha) \), statements we are in favor of.

3. A significant test is a decision rule based on the data which decides whether to reject \( H_0 \).

4. Errors.
   - Type I error: \( \alpha = P(\text{reject } H_0, \text{ when } H_0 \text{ is true}) \)
   - Type II error: \( \beta = P(\text{fail to reject } H_0, \text{ when } H_a \text{ is true}) \)

5. \( \alpha \) level significance test: A test whose type I error is \( \alpha \).

6. The p- value: The smallest \( \alpha \) at which \( H_0 \) can be rejected.
Facts and Formulas

Chapter 7: Confidence Intervals

One Sample Mean:

The basic form of a 100(1-\(\alpha\))% confidence interval for an unknown population mean, \(\mu\), is

\[
\bar{x} \pm C^* \frac{\sigma^*}{\sqrt{n}}, \text{ which is the same as } \left[ \bar{x} - C^* \frac{\sigma^*}{\sqrt{n}}, \bar{x} + C^* \frac{\sigma^*}{\sqrt{n}} \right],
\]

where \(C^*\) comes from the appropriate table and \(\sigma^*\) denotes the appropriate population SD or estimate thereof. The meaning of such an interval is that \(\text{Prob}(\mu \in \text{Interv}) = 1-\alpha\), at least approximately. (The term \(\frac{\sigma^*}{\sqrt{n}}\) corresponds to (an estimate of) the SD of \(\bar{X}\), and hence is sometimes referred to as the SE.) The important situations are:

- \(\sigma\) known and population normal or \(n\) large: \(\sigma^* = \sigma\) and \(C^* = z_{\alpha/2}\), from standard normal tables.
- \(\sigma\) unknown and \(n\) large: \(\sigma^* = S\) and \(C^* = z_{\alpha/2}\), from standard normal tables.
- \(\sigma\) unknown and population normal: \(\sigma^* = S\) and \(C^* = t_{\alpha/2, \text{df}=n-1}\) from the t-table.

The sample size needed to get a total interval width of \(w\) is (approximately)

\[
n = \left\lfloor 2z_{\alpha/2} \frac{\sigma^*}{w} \right\rfloor^2.
\]

If \(\sigma\) is not known, use an estimate from previous experience or from the corresponding value of \(S\) in a pilot experiment (as described in the problem).

One population proportion:

When \(n\) is large and \(\hat{p}\) is not too near 0 or 1 you can use the classical “large-sample” formula \(\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}\). (Large here is approximately \(n\hat{p}\hat{q} \geq 20\).)

Otherwise use the “score” ("Wilson’s") formula

\[
\hat{p} + \frac{z^2}{2n} \pm z \sqrt{\frac{\hat{p}\hat{q}}{n} + \frac{z^2}{4n^2}} \quad \text{with } z = z_{\alpha/2}.
\]

The sample size needed to get a total interval width of \(w\) is (approximately)

\[
n = \left\lfloor 2z_{\alpha/2} \frac{\sqrt{\hat{p}\hat{q}}}{w} \right\rfloor^2 \text{ where } \hat{p} \text{ is a prior guess for the true value of } p.
\]
If no good information is available to guess the true value of \( p \) then use \( \tilde{p} = \frac{1}{2} \). This gives a conservative (possibly too large) choice of \( n \). (The formula for \( n \) is based on the large-sample formula. If the \( n \) turns out small, this will not be a very accurate answer, but use it anyway.)

**Prediction Intervals:**

When the population is normal a prediction interval for a single future observation takes the form \( \bar{x} \pm C^* \cdot \sigma^* \sqrt{1 + \frac{1}{n}} \) with \( C^* \) and \( \sigma^* \) as above. If we denote that future observation by \( Y \) such an interval has the property that \( \text{Prob}(Y \in \text{PredInterv}) \approx 1-\alpha \).

**Chapter 8: Hypothesis Tests**

**General Theory**

Tests involve a “null hypothesis”, \( H_0 \), and an “alternative hypothesis”, \( H_a \). A typical case is \( H_0: \mu = \mu_0 \) and either \( H_a: \mu \neq \mu_0 \), (“two-sided”) or \( H_a: \mu < \mu_0 \), (“one-sided”). They involve a “test statistic”, call it \( T \) for now, and a rejection region such as \( |T| > K \) or \( \Upsilon < K \).

Suppose we are testing an unknown mean, \( \mu \). Then
\[
P_{\mu \in H_0}(\text{Reject } H_0) = Pr_{\mu}("\text{Type I Error}") ,
\]
and
\[
P_{\mu \in H_a}(\text{DoNotReject } H_0) = Pr_{\mu}("\text{Type II Error}") = \beta \text{ at this } \mu .
\]

The significance level, \( \alpha \), is the probability of type I error at the boundary value, \( \mu_0 \), that divides \( H_0 \) and \( H_a \). One related term is the “**Power at \( \mu \)** = \( P_{\mu}(\text{Reject } H_0) = 1-\beta \).

If one is testing \( H_0 \) as above and observes a value \( \Upsilon = \tau \) then the “\( P \)-value” corresponding to \( \tau \) is the value of \( \alpha \) for which \( H_0 \) is just barely rejected (or just barely not rejected). If you know \( P \) and \( \alpha \) then you know the outcome of the test since
\[
\alpha > P \Rightarrow \text{Reject } H_0 \quad \text{and} \quad \alpha \leq P \Rightarrow \text{DoNotReject } H_0 .
\]

For the usual two-sided tests a level \( \alpha \) test **DoesnotReject** exactly when \( \mu_0 \) is in the \( 1-\alpha \) confidence interval.

**Particular Tests**

**One sample mean:**

A **test** of \( H_0: \mu = \mu_0 \) vs \( H_a: \mu \neq \mu_0 \) has the form: Reject if \( |\Upsilon| > C^* \) where
\[
\Upsilon = \frac{\bar{X} - \mu_0}{\sigma^*} \sqrt{\frac{n}{\sqrt{n}}}
\]
and $\sigma^*$ and $C^*$ are as in the above discussion of two-sided confidence intervals.

The **P-value** corresponding to $\Upsilon = \tau$ can be found from a normal or t-table (depending whether $n$ is large or small) by looking up

$P(|Z| > \tau)$ in the normal case,

and $P(|t_{df=n-1}| > \tau)$ in the t case.

NOTE that because of the $| |$ sign these involve multiplying the tabled values times 2.

The sample size $n$ for which a two–sided level $\alpha$ test of $\mu_0$ has Type II error = $\beta$ at the alternative $\mu'$ is

$$n = \left[ \frac{\sigma(z_{\alpha/2} + z_\beta)}{\mu_0 - \mu'} \right]^2 \text{ (approximately)}$$

In the one sided case put $z_\alpha$ in place of $z_{\alpha/2}$. (The answer should be a large $n$ since this formula involves the normal tables.)

A test of the one sided alternative $H_a: \mu < \mu_0$ would reject when $\Upsilon < C^{**}$ where $C^{**}$ is as in the case of a one sided lower confidence bound. P-values can be found analogously. NOTE they do not involve multiplying tabled entries times 2.

**One population proportion:**

For testing $H_0: p = p_0$ use

$$\Upsilon = \frac{\hat{p} - p_0}{\sqrt{p_0q_0/n}}$$

with rejection regions determined in the usual manner from the standard normal table (as in the case of one sample mean). P-values are also determined from this $\Upsilon$ in an analogous manner. (n should not be too small here – $np_0q_0 > 5$ should suffice. There is a procedure based on the binomial distribution that can be used for smaller $n$.)

The formula for the $n$ for which a two-sided test of $p_0$ also satisfies Type II error = $\beta$ when $p = p'$ is

$$n = \left[ \frac{z_{\alpha/2} \sqrt{p_0q_0} + z_\beta \sqrt{p'q'}}{p' - p_0} \right]^2 \text{ (approximately)}.$$  

For a one-sided test substitute $z_\alpha$ for $z_{\alpha/2}$.

**Chapter 8: Inferences based on Two-Samples**

**Difference of Two means:**

The null hypothesis will be of the form $H_0: \mu_1 - \mu_2 = \Delta_0$. (Usually $\Delta_0 = 0$.) If the samples from the two means, $\bar{X}$ and $\bar{Y}$, are independent then use
\[ \Upsilon = \frac{\bar{X} - \bar{Y} - \Delta_0}{\sqrt{\frac{(\sigma_1^*)^2}{n_1} + \frac{(\sigma_2^*)^2}{n_2}}} \]

with \( \sigma^* \) and \( C^* \) as in the one sample procedures, above. In the case where \( n_1 \) or \( n_2 \) are small and the values of \( \sigma \) are not known in advance you need to assume normality, and then treat \( \Upsilon \) according to a t-distribution with \( \nu \) df, where

\[ \nu = \frac{\left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}} \]

Note that \( \min\{n_1-1, n_2-1\} \leq \nu \leq n_1 + n_2 + 1 \).

A 100(1-\( \alpha \)) % confidence interval for \( \mu_1 - \mu_2 \) takes the form

\[ \bar{X} - \bar{Y} \pm C^* \times SE \]

where SE is the denominator of the test statistic, \( \Upsilon \).

P-values can be found in the usual way from the value of \( \Upsilon \) and the corresponding table.

Formulas for \( \beta \) can be derived from the structure of the test. Sample size calculations are complicated, except for some special cases, and we do not give general formulas here, or for two proportions, below.

If the additional assumption that \( \sigma_1 = \sigma_2 \) seems tenable then use \( s_p \) as the common estimate for \( \sigma \) and \( n_1 + n_2 - 2 \) as the degrees of freedom.

**Means of Paired Samples:** When the data are paired in any possibly meaningful way the null hypothesis \( H_0: \mu_1 - \mu_2 = \Delta_0 \) should be tested by a matched pairs test. This test computes the \( n \) differences \( D_i = X_i - Y_i \) and then uses these for a one-sample test of \( H_0: \mu_D = \Delta_0 \).

**Difference of two proportions:** For testing \( H_0: p_1 = p_2 \) the situation is entirely similar to the preceding, except that we now use

\[ \Upsilon = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \]

along with the normal table, where

\[ \hat{p} \]

is the COMBINED SAMPLE estimate of the proportion, defined by

\[ \hat{p} = \frac{X + Y}{n_1 + n_2} = \frac{n_1}{n_1 + n_2} \hat{p}_1 + \frac{n_2}{n_1 + n_2} \hat{p}_2. \]
(For validity of this test the $n_1$ and $n_2$ should be not too small; similar to what is needed in the case of testing one proportion.)

P-values can be computed in the usual fashion.

**NOTE** that for a Confidence Interval for the difference of two proportions you should not use the SE based on the pooled estimate, $\hat{p}$, above. Instead you should use

$$\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

as the SE.

**Equality of variances:**
Suppose $S_1^2$ and $S_2^2$ are the sample variances from independent samples of size $n_1$ and $n_2$. Assume the populations are (approximately) normally distributed with respective population SDs $\sigma_1$ and $\sigma_2$. To test $H_0: \sigma_1 = \sigma_2$ use the statistic

$$F = \frac{S_1^2}{S_2^2}.$$ 

Refer the values of this statistic to an F-table with $\nu_1 = n_1-1$ df in the numerator and $\nu_2 = n_2 -1$ df in the denominator. Our F table – like most others – is very incomplete, and a little awkward to use. We will usually get the F-values from JMP. (On an exam suitable values from JMP – and maybe some unsuitable ones – will be separately supplied.) For a 2-sided test of this $H_0$ reject if $F$ is either too small or too large.