Model Based Statistics in Biology.

Part II. Quantifying Uncertainty.

Chapter 7.3 Hypothesis Testing with Distribution Functions

ReCap. Part I (Chapters 1,2,3,4)
ReCap Part II (Ch 5, 6)
7.0 Inferential Statistics
7.1 The Logic of Hypothesis Testing
Rejecting the 'Just Luck' Hypothesis
Three Styles of Inference
The Logic of the Null Hypothesis
Choice of Alternative Hypotheses
Type I and Type II Error
7.2 Hypothesis Testing with an Empirical Distribution
7.3 Hypothesis Testing with Cumulative Distribution Functions
7.4 Parameter Estimates
7.5 Confidence Limits

```
For each example,
draw graph of cdf
Show one arrow up and across for
one-tailed test
Show two arrows up and across for
two-tailed test
```

on chalk board

ReCap (Ch 6)
Frequency distributions are a key concept in statistics.
They are used to quantify uncertainty.
Empirical distributions are constructed from data
Theoretical distributions are models of data.

ReCap (Ch 7)
Inferential statistics are a logical procedure for making decisions when there is uncertainty due to variable outcomes. Frequentist decision making is based on the logic of rejecting the null (Just Luck) hypothesis. p-values are calculated from the distribution of outcomes when the null hypothesis is true. P-values can be calculated from empirical distributions obtained by randomizing the data.

Today: More examples, using a generic recipe for statistical inference.
This time with a cumulative distribution function to compute a p-value

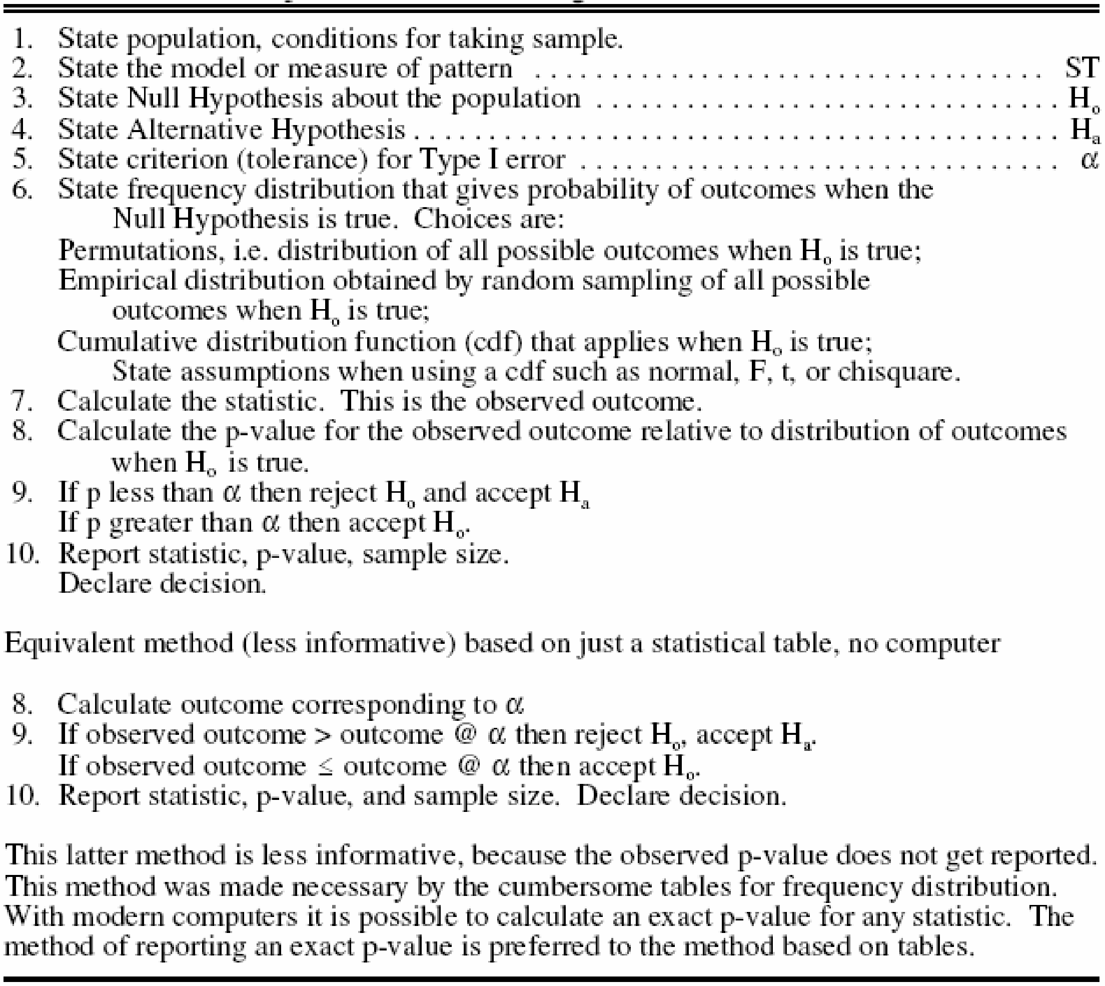
Wrap-up

Hypothesis testing--a set of rules for making decisions in the face of uncertainty. Example of Izaak Walton: Skill or JUST LUCK ?

H_o: Data = Noise (no signal)
H_A: Data = Signal + Noise
Type I error: accepting H_A when in fact H_o is true. This is the *p*-value.
vs Type II error: rejecting H_A when in fact H_A is true

p-value was computed from cumulative distribution function.

Table 7.1 Generic recipe for decision making with statistics.



1. State population, conditions for taking sample.
2. State the model or measure of pattern .... ST
3. State Null Hypothesis about the population .... H_o
4. State Alternative Hypothesis .... H_a
5. State criterion (tolerance) for Type I error .... α
6. State frequency distribution that gives probability of outcomes when the Null Hypothesis is true. Choices are:
   Permutations, i.e. distribution of all possible outcomes when H_o is true;
   Empirical distribution obtained by random sampling of all possible outcomes when H_o is true;
   Cumulative distribution function (cdf) that applies when H_o is true;
   State assumptions when using a cdf such as normal, F, t, or chisquare.
7. Calculate the statistic. This is the observed outcome.
8. Calculate the p-value for the observed outcome relative to distribution of outcomes when H_o is true.
9. If p less than α then reject H_o and accept H_a
   If p greater than α then accept H_o.
10. Report statistic, p-value, sample size.
    Declare decision.

Equivalent method (less informative) based on just a statistical table, no computer

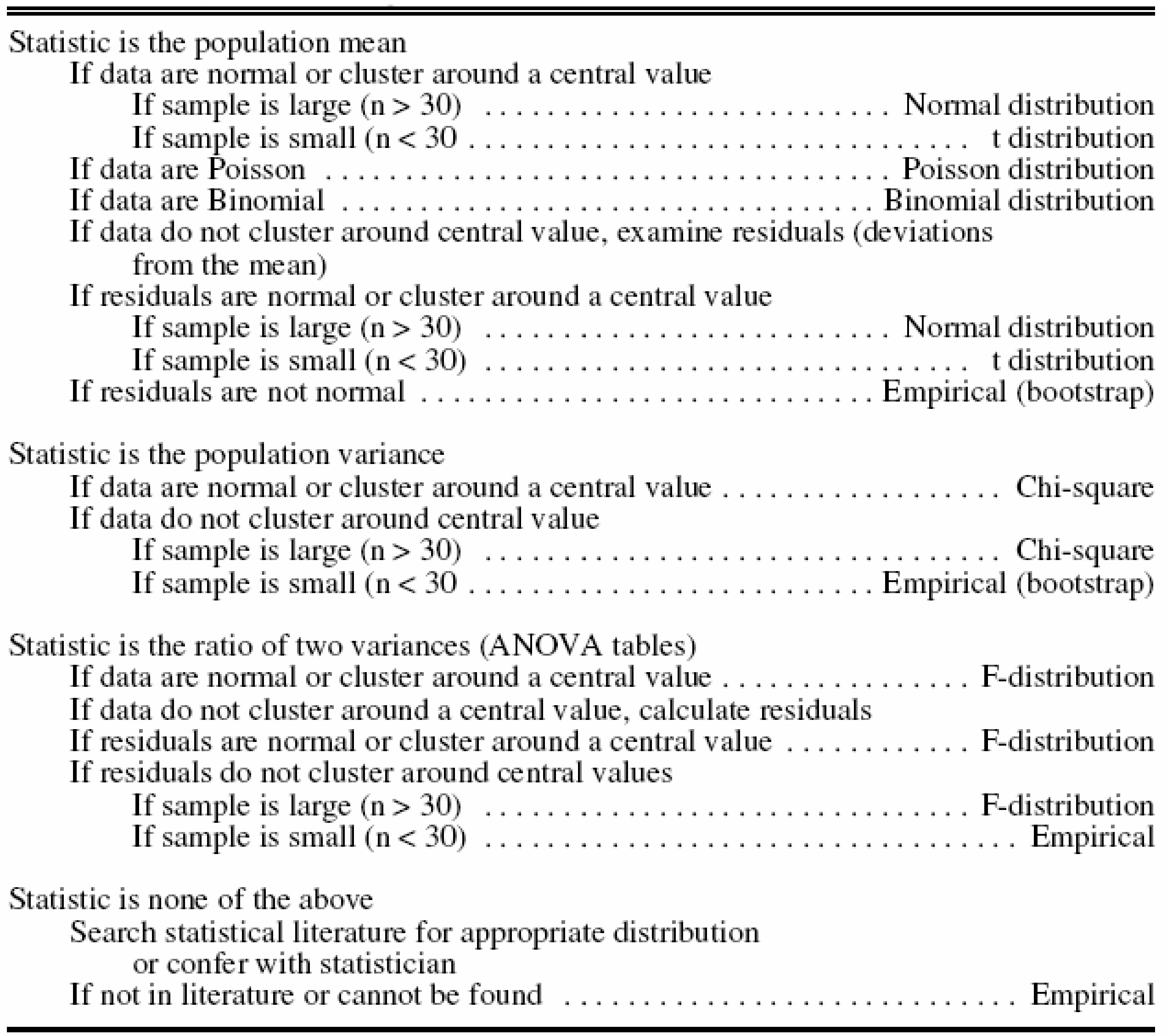
8. Calculate outcome corresponding to α
9. If observed outcome > outcome @ α then reject H_o, accept H_a.
   If observed outcome $\leq$ outcome @ α then accept H_o.
10. Report statistic, p-value, and sample size. Declare decision.

This latter method is less informative, because the observed p-value does not get reported. This method was made necessary by the cumbersome tables for frequency distribution. With modern computers it is possible to calculate an exact p-value for any statistic. The method of reporting an exact p-value is preferred to the method based on tables.

Fisher's famous paper of 1922, which quantified information almost half a century ago, may be taken as the fountainhead from which developed a flow of statistical papers, soon to become a flood. This flood, as most floods, contains flotsam much of which, unfortunately, has come to rest in many text books. Everyone will have his own pet assortment of flotsam; mine include most of the theory of significance testing, including multiple comparison tests, and non parametric statistics.

John Nelder, Rothamsted Experimental Station. (Fisher's successor as Director of the Statistics Department, and pioneer of generalised linear models). From: *Mathematical Models in Ecology*, British Ecological Society Symposium 1971.

Table 7.2. Key for choosing the frequency distribution of a statistic.



Statistic is the population mean	
If data are normal or cluster around a central value	
If sample is large (n > 30)	Normal distribution
If sample is small (n < 30	t distribution
If data are Poisson	Poisson distribution
If data are Binomial	Binomial distribution
If data do not cluster around central value, examine residuals (deviations from the mean)	
If residuals are normal or cluster around a central value	
If sample is large (n > 30)	Normal distribution
If sample is small (n < 30)	t distribution
If residuals are not normal	Empirical (bootstrap)
Statistic is the population variance	
If data are normal or cluster around a central value	Chi-square
If data do not cluster around central value	
If sample is large (n > 30)	Chi-square
If sample is small (n < 30	Empirical (bootstrap)
Statistic is the ratio of two variances (ANOVA tables)	
If data are normal or cluster around a central value	F-distribution
If data do not cluster around a central value, calculate residuals	
If residuals are normal or cluster around a central value	F-distribution
If residuals do not cluster around central values	
If sample is large (n > 30)	F-distribution
If sample is small (n < 30)	Empirical
Statistic is none of the above	
Search statistical literature for appropriate distribution or confer with statistician	
If not in literature or cannot be found	Empirical

Empirical distributions are generated by taking all permutations, by sampling permutations, or by subsampling (bootstrap methods).

Examples of Hypothesis Testing, Using a Generic Recipe
Jackal bones again

Data from Manly (1991) analyzed again with same generic recipe, but this time with different ingredients, to show that there is more than one "correct" analysis of a data situation.

1. State population.
   Unknown. Set of 20 bones in a museum.
   We do not know whether these are representative of jackal populations
      This is typical of text examples.
      Going to take population as all possible measurements on these bones, rather than all jackals in the world.
   Thus we are looking at measurement error, not process error (due to biological processes).

2. Measure of pattern.
   The statistic used in the previous analysis was the difference in mean lengths.
   $D_o \quad = \quad L♀ \quad - \quad L♂ \quad = \; -4.8$ mm
   In this example, the statistic will be a standardized difference,
   called the t-statistic.

$$St = \; t = \frac{D_o}{\sqrt{\frac{1}{n}\left(\text{var}(L_{male}) + \; \text{var}(L_{female)}\right)}}$$

n = 10 bones, in each of two samples
(formula is for equal sample sizes in two groups)

Var(L♀)= variance of 10 measurements of L♀
Var(L♂)= variance of 10 measurements of L♂
D_o is difference in mean lengths, the same statistic that was used in the previous analysis of this data.

the t-statistic is a difference, standardized by the square root of a variance.
the t-statistic is a ratio of two quantities with the same units
   In the jackal bone example the units of t are mm/mm.
The t-statistic has no dimensions or units, it is just as useful for mice as for microbes or megatheria.

Here is the general formula for the t-statistic

$$t = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) s_p^{\,2}}}$$

$\overline{X}_1$ $\overline{X}_2$ are mean values of X in samples from populations 1 and 2

μ_1 μ_2 are the true means in populations 1 and 2

n_1 n_2 are sample sizes from populations 1 and 2

$s_p^{\,2}$ is the pooled variance

$$s_p^{\,2} = \frac{(n_1 - 1)s_1^{\,2} + (n_2 - 1)s_2^{\,2}}{n_1 + n_2 - 2}$$

$s_1^{\,2}$ $s_2^{\,2}$ are the variances in samples 1 and 2

When the null hypothesis is true, $(\mu_1 - \mu_2) = 0$

$$t = \frac{\left(\overline{X}_1 - \overline{X}_2\right)}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) s_p^{\,2}}}$$

When sample sizes are equal, the formula becomes

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{1}{n}\left(s_1^{\,2} + s_2^{\,2}\right)}}$$

Hypothesis testing -- Jackal bones (continued)

3. H_A: $t \neq 0$ (i.e., $L♀ \neq L♂$)
4. H_o: $t = 0$ (note: this is two-tailed test, unlike previous analysis)
   Example demonstrates *p*-value from cdf, for a two-tailed test.

5. $\alpha = 5\%$ We will tolerate Type I error of 5%, *i.e.,* error of accepting a difference that does not exist

6. Work through key, to *t*-distribution.

7. Calculate the t-statistic

$$t = \frac{113.4 - 108.6}{\sqrt{\frac{1}{10}(13.82 + 5.16)}}$$

$t = 3.484$

8. Calculate Type I error (p-value) from *t* distribution

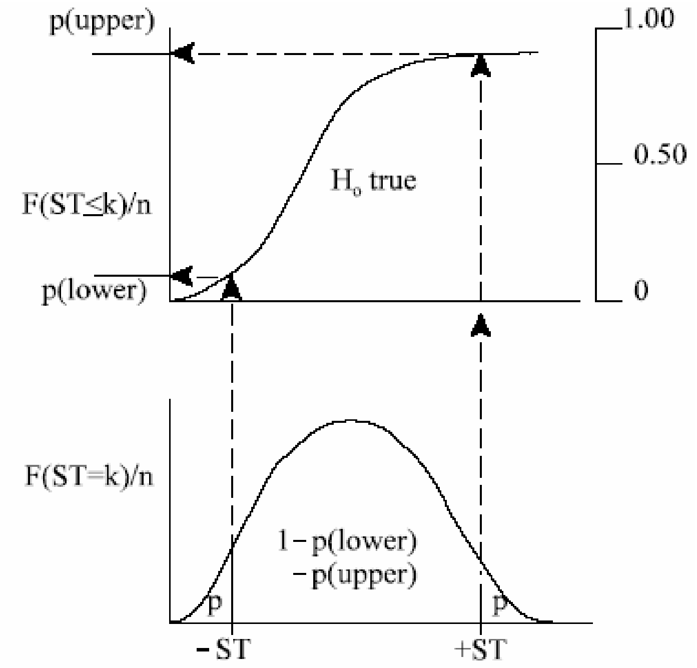
p can be calculated for the observed $t = 3.484$ with 18 df in any package

```
MTB > cdf   3.484;
SUBC> t 18.
           3.484    0.9987
```

18 degrees of freedom = 20 − 2
lose 1 df for each parameter estimated
two parameters (means) were estimated

+3.484	0.9987	$p = 1 - \text{cdf} = 1 - 0.9987 = 0.0013$
−3.484	0.0013	$p = 0.0013$

Figure L11a. Show both lines coming across, one for each tail.



Hypothesis testing -- Jackal bones (continued)

8. Calculate *p*-value from *t* distribution
   repeat for lower tail because this is a two tail test, that bones from males are
   either longer or shorter than from females.

   Add tails together: *p*-value = 0.0013 + 0.0013 = 0.0026
   Compare this to the result from randomization $p = 9/5000 = 0.0018$
   In this case it turns out that the distribution function gives nearly the same p-value as the empirical distribution. The value from a mathematical distribution function is quicker to calculate than the p-value from an empirical distribution generated by Monte Carlo methods.

9. $p < \alpha$ so reject H_o ($t = 0$, hence $D_o = 0$) in favour of H_A ($D_o > 0$)

10. $t = 3.484$, $n = 20$, $p = 0.0016$
    male jackal bones significantly longer than female bones.

    or: *L*♀ = 108.6 *L*♂ = 113.4 $t = 3.484$ df = 18 $p = 0.0026$

We have eliminated measurement error as an explanation for the observed difference in average length. We have not eliminated process error due to natural variation in mandible lengths among populations, process error due to biased sampling, etc. To do this we would need an unbiased sample from several natural populations.

Equivalent procedure for steps 8, 9, 10 (less informative)

8. *t*-statistic corresponding to $\alpha = 5\%$ is 1.734 for one-tail
   it is 2.10 for two tail test.

9. $t_{\alpha=5\%} = 1.734 < t_{obs} = 3.484$ so reject H_o

10. $t = 3.484$ $p < 0.05$ $n = 20$

This procedure is less informative, and no longer necessary, now that we have computers with easily used and accurate software. This procedure is a carry-over from the days before hand held calculators and personal computers, when p-values had to be tabled rather than calculated exactly. Hand held calculators with programs that calculate *p*-values for *t, F,* and Chisquare distributions appeared in the late 1970s. These days, Type I error can be calculated in spreadsheets, in any statistical package, and from programs on the World Wide Web. Using Tables is like carrying around a rotary dial phone instead of a cellphone.

Hypothesis testing – Checking assumptions.

The *t*-distribution is a theoretical distribution calculated from a mathematical expression.

This distribution applies to the t-statistic we have calculated, provided the deviations from the two means *L*♂ and *L*♀ have a normal distribution.

Note that the assumption is that the residuals (the deviations from the two means) are normal. We cannot check the assumption before computing the two means.

Data = Model + Residual. The model is that of two means.

It is a logical contradiction to check this assumption before undertaking the test: after all, we are expecting bone lengths to differ between females and males, and hence we expect that the data itself (all 20 observations) will be somewhat bimodal (not normal).

Even now, well into the 21st century, you may well encounter someone who insists that you check whether your "data must be normal" before doing the *t*-test. This is not correct. It is a waste of time.

To check assumptions for a *t*-test we examine the histogram of residuals after the parameters (two means in this case) are estimated.

```
Show histogram for females,
males, and both combined.
```

Hypothesis testing -- Roach Survival

Here is another analysis, using the generic recipe for hypothesis testing.
The example is roach survival, from Box 8.1 in Sokal and Rohlf 2012, p187.

Willis and Lewis (1957 *Journal of Economic Entomology* 50: 438-440) investigated the survival time of cockroaches, which disperse in shipping containers in the absence of food and water. The survival of the roach *Blatella vaga* was significantly greater in females than in males, when kept without food or water ($t = 5.52$, $n = 20$, $p < 0.001$). The t-test assumes no difference in variance in survival. Sokal and Rohlf tested whether the variance in survival differed between male and female *B. vaga.*
Survival (T_s) in days of the roach *Blatella vaga* when kept without food or water.

Females	$n = 10$	mean(T_s) = 8.5 days	sterr(T_s) = 0.6 days	var(T_s) = 3.6
Males	$n = 10$	mean(T$_s$) = 4.8 days	sterr(T_s) = 0.3 days	var(T_s) = 0.9

1. Population. Sample is set of 20 measurements. Based on an implicit assumption (population is all roaches of this species) Willis and Lewis concluded that mean survival of female *B vaga* exceeds mean survival of male *B. vaga*.

2. ST = F = var(T_{s_Female})/var(T_{s_Male})
The F-statistic, like the *t*-statistic, depends on the sample size.
It depends not only on the sample size of denominator variance
(as with the *t*-statistic)
it depends also on the sample size of the numerator variance.
The notation will be $F_{\text{df numerator, df denomator}}$
Thus $F_{9,9}$ for $n = 10$ in numerator and $n = 10$ in denominator.

Can we conclude that male and female roaches differ in variance in survival?

3. H$_o$: $F = 1$ i.e. var(T_{s_Female}) = var(T_{s_Male})

4. H$_A$: $F \neq 1$ i.e., var(T_{s_Female}) $\neq$ var(T_{s_Male})
This is a two-tailed test. We have no reason to expect male and female cockroaches to differ in variance in survival.

5. $\alpha = 5\%$

6. We use the F-distribution for the F-statistic.

7. $F_{9,9}$ = var(T_{s_Female})/var(T_{s_Male}) = 4.0
$F_{9,9}$ = var(T_{s_Male})/var(T_{s_Female}) = 0.9/3.6 = 0.25

Hypothesis testing -- Roach Survival (continued)

8. In order to compute the p-value from the F-distribution we need to state degrees of freedom for the numerator and denominator variances.
$F_{9,9} = 4.0$ $p = 0.0254$ upper tail of F-distribution with df = 9,9
$F_{9,9} = 0.25$ $p = 0.0255$ lower tail of F-distribution with df = 9,9
Sum p = 0.0509

9. $p = 5.09\% > \alpha = 5\%$ so accept H_o

10. Variance in survival does not differ between male and female roaches
($F_{9,9,} = 4.0$ $p = 5.09\%$)

Based on an implicit assumption (population is all cockroaches of this species) Willis and Lewis concluded that male and female cockroaches differ in average survival. They further concluded that cockroaches can disperse by shipping almost anywhere in the world.
Sokal and Rohlf (2012) concluded that the assumption of no difference in variance was warranted for the test of difference in means. The substantial difference in variance (4 times higher in female than male *B. vaga*) can be attributed to chance, not to some biological basis for greater variability in survival of females, such as unusual longevity in a small proportion of females, compared to males.

Extensions.

1. Confirm the calculation of $t = 5.52$ for the t-test of means.
2. Given the reported standard deviations, would the conclusion of no difference in variance hold for a sample size of 20 males and 20 females?