Model Based Statistics in Biology.

Part II. Quantifying Uncertainty and Evidence.

Chapter 7.3 Hypothesis Testing with Distribution Functions

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

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)

Frequentist hypothesis testing 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.

Wrap-up

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

Hypothesis testing

H_o: Data = Noise (no signal)

H_A: Data = Signal + Noise

Type I error: rejecting H_o when H_o is true. This is the *p*-value.

Type II error: not rejecting H_o when H_o is false

p-value was computed from cumulative distribution function.

Table 7.1 Generic recipe for frequentist hypothesis testing

1. State background and research question.
2. Define population, sample, and relation of sample to population.
3. State the test statistic..…….……*ST*
4. State null hypothesis about the population.…… H_o
   State research hypothesis about population…… H_A
5. Type I error fixed or categorical?
6. State frequency distribution that gives probability of outcomes when the null hypothesis is true. Choices:
   a) All possible outcomes (permutation test)
   b) Empirical distribution obtained by random sampling of all possible outcomes when H_o is true (randomization test).
   c) Cumulative distribution function (cdf) that applies when H_o is true
      State assumptions when using a cdf such as Normal, F, t or χ^2
7. Calculate the statistic from the sample.
   This is the observed outcome for a randomization test
8. Calculate the p-value for the observed outcome relative to the distribution of outcomes when H_o is true.
9. Reject H_o if p less than α, declare decision about H_o
   OR Evaluate H_o from ranking of p.
10. Report test statistic, p-value, sample size. Report parameter estimates as appropriate. Report measure of evidence (LR) if appropriate.
    Draw science conclusions.

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).

Hypothesis testing with a probability model
Jackal bones again
Data from Manly (1991) analyzed again with same generic recipe, but this time with a probability distribution (*t*-distribution) instead of an empirical distribution.

1. Background and research question.
   No information on research questions provided in Manly (1991)
2. Population: all possible measurements on these bones.
   Thus we are looking at measurement error, not process error (due to other causes). Sample: 20 bones. Representative? Unknown
3. Test statistic.
   The statistic used in the previous analysis was the difference in mean lengths.
   $D_o = L_{female} - L_{male}$ –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
   This formula is for equal sample sizes in two groups.

   $Var(L_{female})$= variance of 10 measurements of L_{female}
   $Var(L_{male})$= variance of 10 measurements of L_{male}
   D_o is difference in mean lengths, the same statistic that was used in the analysis of this data using empirical distribution.

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. Its magnitude does not depend on units of measurement it is just as useful for mice as for microbes or mammoths.
The *t*-statistic puts the data on a new scale with units of standard deviations. The difference measured in mm is transformed to a distance measure in standard deviations.

Hypothesis Testing with a probability model. Jackal bones again

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

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

$s_1^{\,2}$ $s_2^{\,2}$ 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

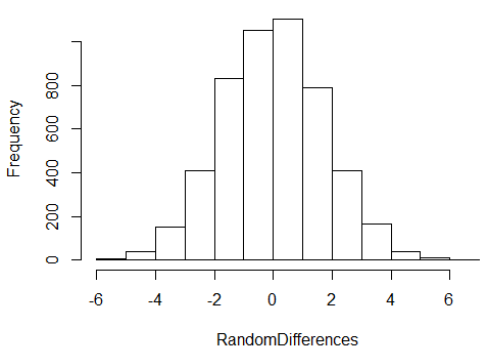
$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{1}{n}\left(s_1^{\,2} + s_2^{\,2}\right)}}$$

Hypothesis testing -- Jackal bones (continued)

4. H_o: $t = 0$ I.e., males and females the same. This is a two-tailed test
   H_A: $t \neq 0$ I.e., males and females not the same
   OR
   H_A: $t < 0$ I.e., males larger. This is a one-tailed test
   H_o: $t \geq 0$ I.e., males not larger

5. Type I error fixed or categorical? We have no reason to control Type I error. Instead we will use 3 categories
   p > 10% : High Type I error
   10% >p>5%: Moderate Type I error
   p ≤ 5%: Low Type I error

6. In the previous example we constructed a distribution by randomization. This makes no assumptions about the distribution of our statistic. But it requires time and care to set up the test.



The distribution looked normal.
This suggests that we use a normal distribution instead of generating our own.
It turns out the normal distribution is for an infinite number of samples. So instead we will use the t-distribution, which is the normal distribution for a limited number of samples.

This has several advantages It saves time and effort
It allows us to infer beyond our sample.

It assumes that the normal (or t) distribution is appropriate for our sample.
We can check this by comparing our distribution to a normal distribution.
We can also justify the assumption by using the law of large numbers, that repeating the measurements an infinite number of times will produce an estimate of the true value of the difference in means.

How good is the assumption?
In this case the fit looks good, by eye.
And we expect an infinite number of repeats with these bones will produce the true value. We might also expect that our measurement protocol, if repeated an infinite number of times, would produce the true difference for all jackals of this species.

We will assume that the data are normally distributed around the two means, with the same variance. With this assumption, we will use the *t*-distribution.

7. Calculate the t-statistic $$t = \frac{113.4 - 108.6}{\sqrt{\frac{1}{10}(13.82 + 5.16)}}$$

$t = 3.484$ for $L_{male} - L_{female}$ $t = -3.484$ for $L_{female} - L_{male}$

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

18 degrees of freedom = 20 – 2

We lose 1 df for each parameter estimated

Two parameters (means) were estimated

R-code

```
> pt(-3.484,18)
[1] 0.001324587
```

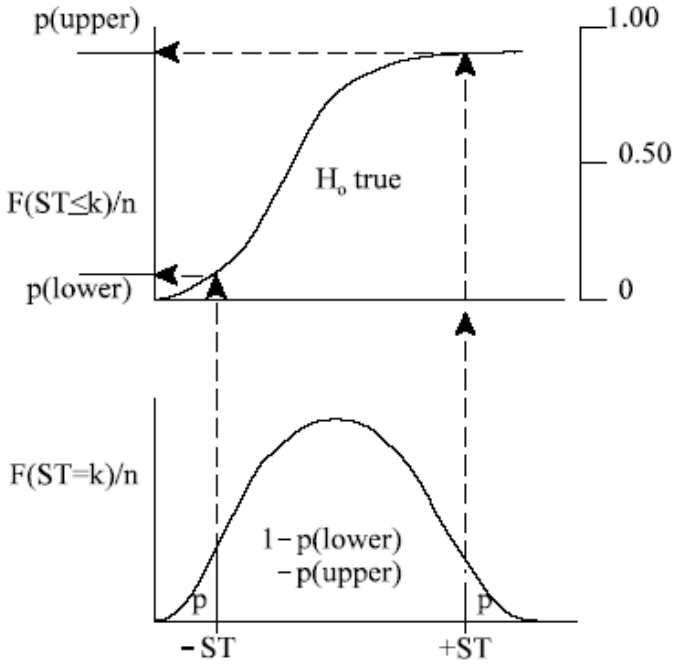
```
MTB > cdf    -3.484;
SUBC> t 18.
          -3.484     0.0013
```

Because t is negative, the functions report the lower tail.

For a one-tail test, we use the lower tail only

The distribution is symmetrical. The upper tail has the same value.

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



Hypothesis testing -- Jackal bones (continued)

8. Compare this to the result from a randomization with more runs.

p_{cdf}	= 0.0013	cumulative distribution function
p_{random} =	9/5000 = 0.0018	previous randomization
p_{random} =	47/20000 = 0.00235	another randomization with greater precision

In this case it turns out that the distribution function gives a smaller p-value.

p_{cdf} : Quicker to calculate.
Assumes that differences are normally distributed.
Inference to an infinite number of repeats.

p_{random} Takes longer to calculate.
Does not assume normal errors.
Does not assume data are representative
Inference only to all rearrangements of the data.

9. Evaluate H_o from ranking of p.
$p < \alpha$ so we discard the H_o (males not larger) at a low (5%) error rate.

10. Report statistics: t = -3.484, n = 20, p = 0.0013 (t-distribution)
We report effect size. L_{female} = 108.6 mm, L_{male} = 113.4 mm

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 a population to infer to that population. We need samples from several populations to infer to the species.

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

8. t-statistic corresponding to α = 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 discard H_o at low (5%) error rate

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 spreadsheet, in any statistical package, and from programs on the World Wide Web. Using Tables is like using a rotary dial phone on your desk, instead of a cellphone.

Hypothesis testing – Checking assumptions for the *t*-distribution.

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_{female} and L_{male} have a normal distribution. If both both groups have normally distributed residuals, then we can assume that the difference of the means is normally distribution. Note that the assumption is about the residuals. We cannot check the assumption until we have the parameter estimates—the mean for each group.
>
> Data = Model + Residual. Here, 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 are normal" before doing the *t*-test. This is not correct. It is a waste of time because the assumption is about the residuals.
Later, when we use the general linear model to carry out a *t*-test we will 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.

1. Research context and question . 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 management context is that quarantines to prevent spread of roaches be set to the survival of females. Setting quarantines depend on the variance in survival – what percent survive at any given quarantine period? 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

2. Define population, sample, and relation of sample to population. Sample is set of 20 measurements. Population is an infinite number of repeats of the experimental protocol in Willis and Lewis. These authors implicitly assumed the results could be inferred to all roaches of this species. Willis and Lewis concluded that mean survival of female *B vaga* exceeds mean survival of male *B. vaga*, a result that could be applied in a pest management context.

3. The test statistic is *F,* the ratio of $\text{var}(T_{s\_\text{Female}})/\text{var}(T_{s\_\text{Male}})$
The F-distribution, like the *t*-distribution, depends on the sample size.
It depends on the sample size of denominator variance
(as with the *t*-statistic) and on the sample size of the numerator variance.
The notation will show degrees of freedom as subscripts: $F_{\text{df numerator, df denomator}}$
Thus $F_{9,9}$ for $n = 10$ in numerator and $n = 10$ in denominator.
Type I error fixed (decision-theoretic) or categories ?
The variance in survival can be used to set isolation times in quarantine
to control spread of roach infested cargo. If variance is greater in females,
a longer quarantine time is needed than using the variance for males and females.
We can identify a risk of not controlling Type I error, so we set it at a
predetermined rate of $\alpha = 5\%$.

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

4. The state of science knowledge is H_o: $F = 1$ *i.e.* $\text{var}(T_{s\_\text{Female}}) = \text{var}(T_{s\_\text{Male}})$
H_A: $F \neq 1$ i.e., $\text{var}(T_{s\_\text{Female}}) \neq \text{var}(T_{s\_\text{Male}})$
This is a two-tailed test. Despite the obvious difference in the standard deviations, we have no reason to expect male and female cockroaches to differ in variance in survival.

Hypothesis testing -- Roach Survival (continued)

5. We have reason to fix Type I error. We set it at 5%

6. We use the F-distribution for the F-statistic. This assumes normal distribution around the fitted values, the two means. To check the assumption we could do a histogram of each of the 20 values, as deviations from their mean.

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

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. 10% > 5.09% > 5% The test has moderate Type I error.
   However, we set the error rate at 5%. So to be consistent we cannot reject the null hypothesis. This example illustrates one of the limitations of using a fixed (Neyman-Pearson decision-theoretic) error rate.

10. $F_{9,9,}$ = 4.0 p = 0.0509 n = 20.
    Statistical conclusion. Even though we found some evidence of a difference in variance in survival of female and male roaches at moderate Type I error of 5-10% we cannot reject the null hypothesis.
    Sokal and Rohlf (2012) used a fixed Type I error of 5%. They concluded that the hypothesis of no difference in variance in survival could not be rejected. However, the sample size was small, and the null hypothesis could potentially have been rejected with a larger sample size.
    Science conclusion. 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.

Extensions.

1. Confirm the calculation of t = 5.52 for the t-test of means.
2. Was the observed difference in variances great enough to reach the 5% criterion with a larger sample? Given the reported standard deviations, would the conclusion of no difference in variance hold for a sample size of 20 males and 20 females?