Model Based Statistics in Biology.

Part II. Quantifying Uncertainty and Evidence.

Chapter 7.2 Hypothesis Testing with an Empirical Distribution

Use Oat yields as first example ?

For each example,

draw graph of cumulative histogram.

Show one arrow up and across for one-tailed test

Show two arrows up and across for two-tailed test

ReCap Part I (Chapters 1,2,3,4)
Quantitative reasoning: Example of scallops, which combined models (what is the relation of scallop density to substrate?) with statistics (how certain can we be?)

ReCap (Ch5)
Data equations summarize pattern in data.

ReCap (Ch 6)
Frequency distributions are another 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 eliminating chance as an explanation for an outcome.

Today: Generic recipe for hypothesis testing.

Wrap-up
We used a generic recipe for statistical decision making based on the logic of the null hypothesis.
To calculate a p-value, we used a distribution of outcomes generated by randomizing the data.

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 (state of science) 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.

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.
   If observed outcome $\leq$ outcome @ α then cannot reject H_o.
10. Report statistic, p-value catgories, 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.

The simplest way of understanding quite rigorously, yet without mathematics, what the calculations of the test of significance amount to, is to consider what would happen if our two hundred actual measurements [of stature of Englishmen and Frenchmen] were written on cards, shuffled without regard to nationality, and divided at random into two new groups of a hundred each. Actually, the statistician does not carry out this very simple and very tedious process, but his conclusions have no justification beyond the fact that they agree with those which could have been arrived at by this elementary method."

Fisher R.A. 1936. Journal of the Royal Anthropological Institute of Great Britain and Ireland 66: 57-63.

Example of Hypothesis Testing, Using the Generic Recipe - Jackal Bones

Example is length of bones from 10 male and 10 female jackals.
L = length of mandible (L= mm) of Golden Jackals
Canis aureus from the British Museum.
Data from Manly (1991).

Male	Female	
120	110	
107	111	
110	107	
116	108	
114	110	
111	105	
113	107	
117	106	
114	111	
112	111	
113.4	108.6	mean
13.82	5.16	var

Generic recipe. Set-up = steps 1-6. Execution = steps 7-10.

1. State background and research question.
   No information provided by Manly (1991).
2. Define population, sample, and relation of sample to population.
   It could be taken as all possible measurements on these 20 bones.
   The values would vary because of measurement error.
   It would be very safe to infer to this population.
   It could be all jackals of this species in the world.
   The values would vary because of individual variation, on top of error.
   Inference to this population is risky. We need to know whether this sample was representative of the population (all jackals)
3. State the measure of pattern (test statistic)
   $ST = D_o = \text{mean}(L_{male}) - \text{mean}(Lfe_{male})$
4. State null (state of science) hypothesis about the population.
   Sexual size dimorphism, with males being larger than females, is well documented in canids: coyotes red foxes, and wolves. Socially monogamous species are only weakly dimorphic in skeletal shape and body mass.
   C. aureus is socially monogamous. We expect weakly dimorphic bone sizes.
   H_o: $D_o < 0$ *i.e.*, $L_{male} - L_{female} < 0$ for the population.
   H_A: $D_o \geq 0$ *i.e.* $L_{male}) - L_{female} \geq 0$ for the population
5. Type I error fixed or categorical? We have no reason to control Type I error, such as risk to subjects (patients) or some preference for fewer false positives (Type I error) than false negatives (Type II error). Instead we will use 3 categories
   $p > 10\%$ High Type I error
   $10\% > p > 5\%$ Moderate Type I error
   $p \leq 5\%$ Low Type I error

Hypothesis Testing Using the Generic Recipe - Jackal Bones

5. Type I error fixed or categorical?

   When would we need to consider a fixed Type I error rate?

   We look at consequences for the subjects (jackals), experimenter, and the published literature.

   In this example the subjects are bones from a museum, there are no consequences for live jackals. If we were using live animals, and there was harm or risk, we would need to develop a protocol that uses a fixed α.

   For the experimenter there is no need to control Type I error, if the experimenter reports the Type I error in each study.

   For the literature, the word "significant" becomes detached from the conventional rate, 5% in most of the natural and social sciences. Reporting the error rate addresses this.

   In an exploratory analysis the threshold is often raised to reduce the chance of Type II error, that something will be missed. This is remedied by calling the threshold a screening criterion, and not declaring significance, with the conventional meaning of 5%.

6. State frequency distribution that gives probability of outcomes when the null hypothesis is true. We will use an empirical distribution. It makes no assumptions about the distribution of our statistics. The frequency distribution of the statistic D_o when the null hypothesis H_o is true will be calculated by randomization.

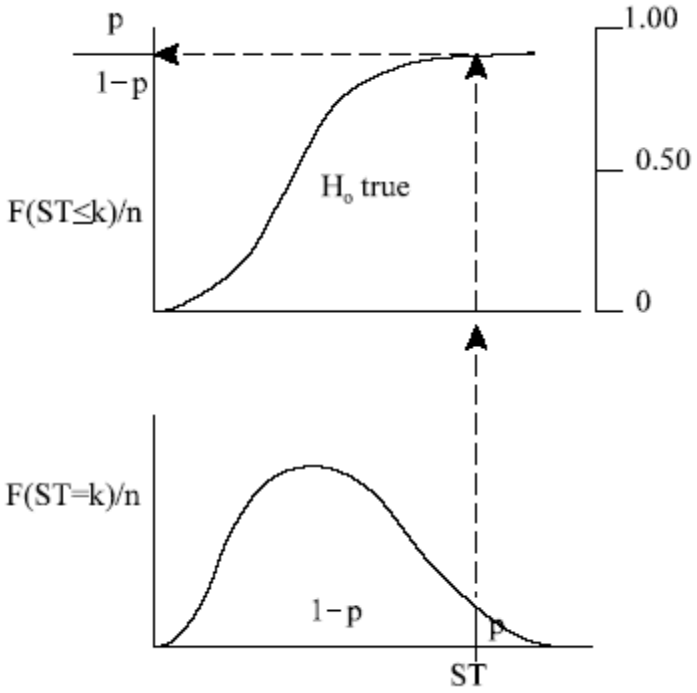
   To obtain the distribution we assign the 20 bones randomly to two groups. To be rigorous, we would use a random number generator to assign bones to two equal sized groups. With a computer we can assign values to groups either with replacement or without.

   Next, we compute the mean for each group. Then we calculate D_o the random difference in means. This is the difference when the null hypothesis is true. We made the null hypothesis true by assigning bones randomly to two groups.

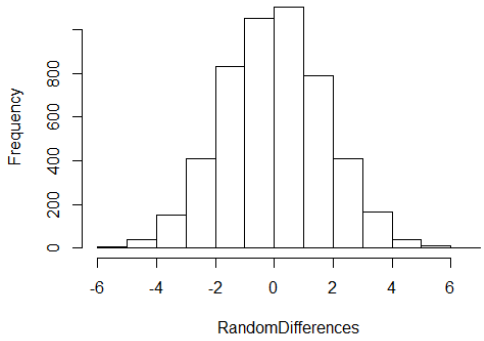
   Next, we repeat this many times to obtain many random differences-- the more the better.

Hypothesis Testing, Using the Generic Recipe - Jackal Bones

6. We assemble these random differences into an empirical frequency distribution.



Figure



7. Calculate the statistic D_o
   from the sample. The observed difference is D_o = 113.4 - 108.6 = 4.8 mm

8. Use the observed difference to estimate Type I error.
   5000 values of D_o Of these 9 values exceed 4.8 p = 9/5000 = 0.18%

9. Evaluate H_o in categories of p.
   We will use Fisher sorting, instead of using a fixed criterion.
   Type I error rate is low according to our categories: $p < 5\%$

10. Report result of test. D_o = 4.8 mm,
    n = 20,
    p = 0.0018
    Parameter estimates are L_{male} = 113.4 mm, L_{female} = 108.6 mm

Compare general procedure (A,B,C,D) with recipe.

A. Define the population (step 1) and the signal (step 2, step 4, step 7)
B. Describe the noise (step 3, step 6)
C. Evaluate signal relative to noise (step 8)
D. Declare a decision (step 9, step 10)

Not used in 1997 onward

Hypothesis testing–Direct versus indirect method.

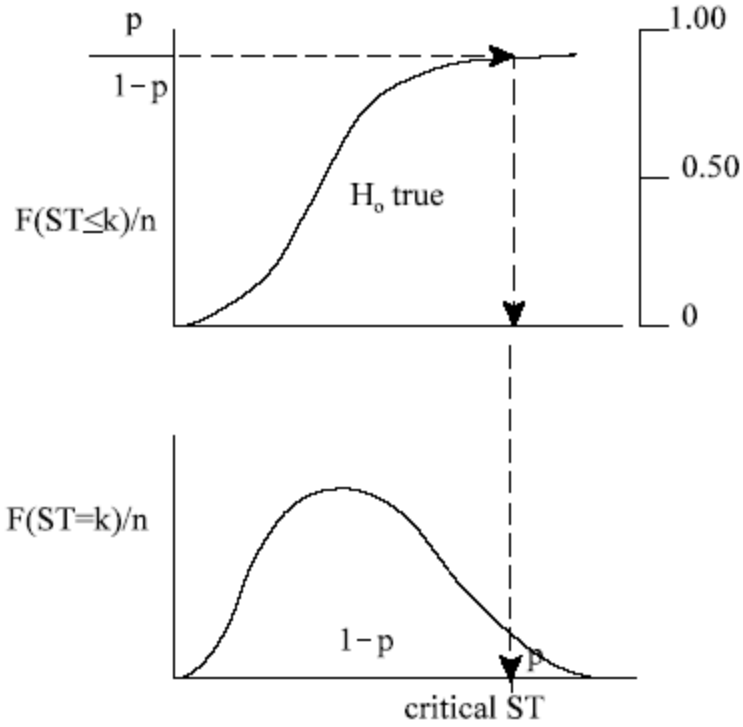
Direct method–compare p-value to criterion.

compute the p-value and compare it with the 5% criterion.
show arrow from statistic, upward to cdf, across to p-value
(Figure L10a on previous page).

Indirect method–compare observed statistic to critical value of statistic.

use criterion (e.g. 5%) to compute critical value corresponding to 5%
show arrow from 1−p across to cdf, downward to critical value.
compare observed value of statistic to this critical value.

Figure L10b



The indirect method is less informative.
It is used when there is no computer available to compute exact p-value.

This course will emphasize the direct method–calculating the p-value.
The direct method provides more information to the reader of a report.
It is consistent with modern practice.
It demonstrates the machinery of hypothesis testing,
which is based on making a decision from a p-value relative to a fixed criterion.
Critical values do this indirectly, rather than directly.

Hypothesis Testing. 2nd example. Heterozygosity data.

TBA. Revision needed to 2019 recipe using Fisher sorting.

Recall material on data equations. Showing improvement (reduction in sum of square deviations)

H_o: $H = \hat{\beta}_o + \epsilon$ $\quad SS_{tot}$ $\Sigma res^2 = \underline{0.1171}$

H_A: $H = \hat{\beta}_E \cdot E + Offset + \epsilon$ $\quad \Sigma res^2 = \underline{0.0204}$

The reduction in squared deviation is: $\quad SS_{model}$ $\Sigma res^2 = \underline{0.0966}$

H = heterozygosity, E = elevation

$H = -0.1273 \cdot E + 0.58 +$ res

Data =	Model	+ residual	
0.1171 =	0.0966	+ 0.0204	LR = $(0.0204 / 0.1171)^{-7/2} = 37$
SS_{tot} =	SS_{model}	+ SS_{error}	

The altitudinal gradient model is more likely than not, LR > 20

The explained variance is $R^2 = 0.096644 / 0.11709 = 83\%$

Is the improvement in fit (SS = 0.0966) better than chance ?

Set up the analysis

1. Sample = 7 measurements.
   Population = all possible measurements taken with a stated procedure.
2. Test statistic = $SS_{model} = 0.0966$, the improvement in fit going from
   $H = \beta_o$ to $H = \beta_o + \beta_E E$ (E = elevation)
3. H_o: $SS_{model} = 0$ (in population) Improvement is 'just chance'
4. H_A: $SS_{model} > 0$ (in population) Improvement is more then 'just chance'
5. $\alpha = 5\%$ (Type I error held to 5%)
6. go to key. SS_{model} not listed, hence use empirical.
7. Execute the analysis. $SS_{model} = 0.096644$ This is the observed improvement.

Calculate frequency distribution of improvements when H_o is true.
Do this by randomizing the heterozygosity data relative to explanatory variable Elev
The estimate parameters and improvement in fit.

Here is R code for to calculate random improvement in fit.

```
```

Here is Minitab code to calculate random improvement in fit.

```
MTB > sample 7 c2 c6
MTB > regress c6 1 c1;
SUBC> residuals c7.
```

Hypothesis Testing. 2nd Example. Improvement in fit, heterozygosity data

TBA. Revision needed to 2019 recipe using Fisher sorting.

Here are the parameter estimates and ANOVA table for a single randomization.

```
The regression equation is
C6 = 0.263 +0.000015 Elev

Predictor       Coef       Stdev    t-ratio        p
Constant      0.2627      0.1184       2.22    0.077
Elev      0.00001506  0.00001787       0.84    0.438

s = 0.1432      R-sq = 12.4%     R-sq(adj) = 0.0%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    1     0.01455     0.01455      0.71    0.438
Error         5     0.10253     0.02051
Total         6     0.11709
```

Random improvement is 0.01455 LR = $(0.10253/0.1171)^{-7/2} = 1.59$

Run this repeatedly with a computer (2000 runs in BrusRN2.out)
[Handout unique randomization to each student]
Then tabulate on chalkboard, to construct frequency distribution.
k = Outcomes(SS_{model}) $F(SS_{model}=k)$

```
MTB > hist c10
Midpoint   Count
    0.00     692  **************************************************
    0.01     422  ******************************
    0.02     283  ********************
    0.03     219  ***************
    0.04     118  ********
    0.05      88  ******
    0.06      69  *****
    0.07      39  ***
    0.08      36  ***
    0.09      16  **
    0.10      16  **
    0.11       2  *
```

Hypothesis Testing. 2nd Example. Improvement in fit, Fly heterozygosity data.

TBA. Revision needed to 2019 recipe using Fisher sorting.

Now compute the number of random improvements that were larger than the observed improvement of 0.096644

```
MTB > hist c10;
SUBC> start 0.096644.

Histogram of C10   N = 2000
1983 Obs. below the first class

Midpoint   Count
   0.097       2  **
   0.098       6  ******
   0.099       0
   0.100       4  ****
   0.101       0
   0.102       3  ***
   0.103       0
   0.104       0
   0.105       0
   0.106       1  *
   0.107       0
   0.108       1  *

MTB > let k1 = 2000-1783
MTB > print k1
K1       17

MTB > let k2 = 17/2000
MTB > print k2
K2       0.00850000
```

8. p = ___/_____ (class result, show of hands to obtain distribution)

Figure 10a, one line coming across cdf frequency distribution, one tailed test.

9. $p = 0.0085 < 0.05 = \alpha$ so reject H_o
The improvement is better than random.

10. $SS_{model} = 0.096644$ $n = 7$ $p = 0.0085$ so reject H_o we reject chance, that is, we reject the JUST LUCK hypothesis.

Hypothesis Testing. 3rd Example. Improvement in fit, Oat Yield data.

TBA. Revision needed to 2019 recipe using Fisher sorting.

```
Recall material on data equations.
Showing improvement (reduction in
sum of square deviations)
```

H_o: $Q = \beta_o$ $\Sigma res^2 = 493.14$ = SStotal

H_A: $Q = \beta_o + \beta_x X$ $\Sigma res^2 = 301.06$

The reduction in squared deviation: $\Sigma res^2 = 192.08$ = SSmodel

Is this improvement better than random ?

Set up the analysis

1. Sample = 8 measurements.
   Population = all possible measurements taken with a stated procedure.
2. ST = SS_{model} the improvement in SS going from $H = \beta_o$ to $H = \beta_o + \beta_E X$ (X = group)
3. H_o: $E(SS_{model}) = 0$ The expected value in the population is zero.
4. H_A: $(SS_{model}) > 0$ The expected value in the population is not zero.
   Note that sums of squares (SS) cannot be less than zero.
5. $\alpha = 5\%$
6. go to key. SS_{model} not listed, hence use empirical (randomization test)

Execute the analysis.

7. $SS_{model} = 192.08$ This is the observed improvement.
   $LR = (301.06 / 493.14)^{-8/2} = 7.2$

Data Equations for null and alternative models

	data	null model	res	res^2	alt. model	res	res^2	observed improvement
0	42.90	40.95	1.95	3.80	36.05	6.85	46.92	
0	41.60	40.95	0.65	0.42	36.05	5.55	30.80	
0	28.90	40.95	-12.05	145.20	36.05	-7.15	51.12	
0	30.80	40.95	-10.15	103.02	36.05	-5.25	27.56	
1	49.50	40.95	8.55	73.10	45.85	3.65	13.32	
1	53.80	40.95	12.85	165.12	45.85	7.95	63.20	
1	40.70	40.95	-0.25	0.06	45.85	-5.15	26.52	
1	39.40	40.95	-1.55	2.40	45.85	-6.45	41.60	
	40.95		0.00	493.14			301.06	192.08

Hypothesis Testing. 3rd Example. Improvement in fit, Oat Yield data.

TBA. Revision needed to 2019 recipe using Fisher sorting

8. Calculate frequency distribution of improvements when H_o is true.
Do this by randomizing the data with respect to explanatory variable X

Data	Equations for null and alternative models							random
	data	null model	res	res^2	alt.model	res	res^2	improvement
0	30.80	42.23	-11.43	130.53	43.70	-12.90	166.41	
0	53.80	42.23	11.58	133.98	43.70	10.10	102.01	
0	49.50	42.23	7.28	52.93	43.70	5.80	33.64	
0	40.70	42.23	-1.53	2.33	43.70	-3.00	9.00	
1	30.80	42.23	-11.43	130.53	40.75	-9.95	99.00	
1	28.90	42.23	-13.33	177.56	40.75	-11.85	140.42	
1	49.50	42.23	7.28	52.93	40.75	8.75	76.56	
1	53.80	42.23	11.58	133.98	40.75	13.05	170.30	
sum			0.00	814.76		0.00	797.35	17.41

Taken from file labeled ST237.xls
Random improvement is 814.76 – 797.35 = 17.41 LR = $(797.35/814.76)^{-8/2} = 1.08$
[note: SStotal now 814 instead of 493 because sampling was with replacement]
[Some values occur twice, some not at all. Mean now 42.23, not 40.95]

8. Calculate *p*-value
Assemble 500 random improvements, compute % that exceed observed improvement of 192.08

count	n	p-value
58	500	0.116

Figure 10a, one line coming across cdf frequency distribution, one tailed test.

9. $p = 0.116 > 0.05 = \alpha$ so accept H_o
The improvement is no better than random.

10. $SS_{model} = 192.08$ $n = 8$ $p = 0.116$ so we H_o not rejected.
that is, the JUST LUCK hypothesis is not rejected.
The sample size is small, which may have been responsible for the decision.

Hypothesis testing. Comparing Variances

TBA. Revision needed to 2019 recipe using Fisher sorting

In some parts of biology, notably population biology the variance is a biologically interpretable statistic. Population biologists (including those who do molecular genetics) think in terms of variances, as much or more than they think in terms of central tendencies measured by means.

For example, balancing selection tends to reduce the variance in a trait, while mutation tends to increase genetic variance and hence increase variance in traits in a population.

Picture here (Fig L13b) of frequency distribution of body size in lizards at time=1 (normal with wide variance), axis labelled $L_{t=1}$ and y-axis labelled $F(L_{t=1})$. Then draw another distribution on another axis labelled $L_{t=2}$ (normal with narrower variance). y-axis labelled $F(L_{t=2})$

Example: Does selection by hawks on young lizards result in balancing selection on body size of lizards ? Size as measured by length L.

Another example: what is the spatial variance in number of species?
What factors tend to reduce species diversity ?

Another example: what is the current level of genetic variability in a population?
What factors tend to increase or reduce genetic variability?

Hypothesis testing. 4th Example. Comparing Variances. Scutum widths

TBA. Revision needed to 2019 recipe using Fisher sorting

Here is another analysis, using the generic recipe for hypothesis testing.
The example is the analysis of scutum widths
Data from Sokal and Rohlf 1995, p808 (2012, p 188)
Hypothesis: Length reflects general genetic variability
Hypothesis: Only ticks from a restricted portion of the genetic spectrum would survive (balancing selection). 25 larval ticks, 16 killed by cold shock, 9 survived. Measurements in mm.

Surviving	Killed
211.3	219.2
211.9	205.1
209.5	213.4
218.5	206.7
204.9	211.1
211.2	222.8
211.4	210.2
205.1	212.7
211.9	210.4
	210.1
	213.1
	224.4
	219.5
	218.4
	204.6
	229.2
210.63	214.43

1. Population: Measurement many times? This would address measurement error, but not the biological hypotheses. As with other experiments, the population can be taken as all possible repeats of the cold shock experiment on this species of tick, given the procedural statement.

2. Test statistic $= F =$ Var(W_{dead})/Var(W_{live})

The biology here is that we expect greater variance in dead than live under most conditions, expect lower survival by individuals with extreme traits than by individuals closer to the mean. That is, expect stabilizing selection under most conditions. Exception is episode of directional selection.

Hypothesis testing. 4th Example. Comparing Variances. Scutum widths

TBA. Revision needed to 2019 recipe using Fisher sorting

3. H_o: $F \leq 1$ i.e. $Var(W_{dead}) \leq Var(W_{live})$
   i.e. no stabilizing selection
4. H_A: $F > 1$ i.e., $Var(W_{dead}) > Var(W_{live})$
   i.e., stabilizing selection
5. $\alpha = 5\%$

6. If we wish to avoid making assumptions about the distribution of the statistic when the null is true, we can generate an empirical distribution to calculate the Type I error (p-value).
   This distribution is generated by randomizing the observations, then repeatedly calculating of the statistic F, to generate a distribution based on many such calculations

7. Calculate F as ratio of two mean squares. That is, $stdev^2$ / $stdev^2$
   In this example we take the numerator as the unreduced variance (killed), taking the denominator as the reduce variance (survived).
   $F = 3.1095$ $LR =$
   If $F < 1$ then accept H_o at this point)
   observed $F > 1$,
   but could this have happened by chance ?
   in other words, what is Type I error ? (if we say F really is > 1)

8. To find out we randomize the data to make the H_o true. Choose 9 ticks randomly, measure, and compare to the remaining. Do this 500 times. The result (S&R2012) was 40 that exceeded F = 3.11
   $p = 40/500 = 8\%$

Figure from Lab 3, showing calculation from distribution where H_o is true.

9. Compare to test criterion. $p = 8\% > 5\% = \alpha$
   H_o: $F \leq 1$ is not rejected

10. Report the result. $F_{15,8} = 3.1095$ $p = 8\%$
    Figure 8.4 in Sokal and Rohlf (2012)
    Note the use of subscripts to show sample size
    The convention is $F_{\text{numerator df, denominator df}}$
    So first subscript is df for W_{dead}

This was a lot of work.
There is an easier way, if we are prepared to make some assumptions.
(Next lecture)