Contributed chapter

Title: Model-based statistical analysis with examples from cancer and stem-cell biology
Author: Michael A. Newton
Version: V2, August 7, 2012
For: Statistics: Discovering Your Future Power
Edited by: Qian Meng et al.

1 Summary

Scientists are able to measure biological systems at incredible levels of detail and also to deploy interventions whose measured effects provide new knowledge about these systems. The variety of systems studied, interventions considered, and effects measured are equally incredible, and one might be surprised to know that a single intellectual framework provides the scientist a guide for how to infer new biological knowledge from the measured data. What we know is that the particular measurements (their numerical values as arranged in whatever Excel spreadsheet, database, or cloud-computing system), while in a way central, are in fact secondary during the intellectual generaton of new insights. Rather, what is primary is the process by which these data have emerged. The numbers are meaningless when stripped of information about their context and details of their generation. In statistical analysis, we achieve understanding by viewing our particular measurements as the realization of a random process. Of course, this is not to say that data are completely unpredictable. But if they were completely predictable, then we would not have bothered to do the necessary experiments to generate them! Indeed, statistics thrives in the grey zone between perfect knowledge and complete uncertainty. Research in this area rewards those who enjoy mathematical and computational thinking, who recognize the primacy of the biological problems, and who are keen to address these problems in order to advance the state of knowledge.

Depending on the nature of the data and the type of system in focus, the precise way in which statistical analyses are deployed and the role of the calculations also vary greatly. In many cases, analysis centers on the construction, use, and assessment of probability models, and a great deal of current research in statistics aims to improve these activities. Very often the investigator has indirect measurements of the objects of primary scientific interest, and the statistical analysis seeks to infer plausible properties of these objects by accounting somehow for signal and noise characteristics of the system and measurement process. This situation is often addressed by constructing probability models of the data and by developing statistical methods based on these models. Two recent examples from my own collaborative work are discussed in the following sections. Statistical concepts employed include maximum likelihood estimation, hypothesis testing, bootstrap sampling, latent variables, nonparametric modeling, and mixture modeling, among others. The examples aim to shed some light on the complex interplay among data context, mathematical analysis, statistical inference, and computation that purvades the modern development of statistics.

2 Inactivating a tumor suppressor gene

Basic biology tells us that some information in DNA is transcribed into messenger RNA and then translated into protein. When the DNA at a specific locus takes two distinct forms, the cell is heterozygous at that locus; indeed when the organism has inherited such different forms (alleles) from its parents, one expects the diploid cells in its body to be heterozygous at that locus. In measuring DNA from many cells of the organism, the two distinct alleles ought to be in 1:1 proportion, though somatic changes may alter the DNA of cancer cells, for example, possibly removing one or the other allele at some locus. Even if the DNA remains intact, the relative abundance of the transcribed RNAs may not be the same for both alleles. For example, epigenetic factors related to accessibility of DNA could cause RNA to be transcribed from only one of the two DNA alleles. I use the term *allelic ratio* to refer to the abundance of one allele relative to another, either in the DNA or RNA, depending on the context.

A chunk of DNA of particular interest in cancer research is the adenometous polyposis coli (APC) gene, which codes for an important protein involved in mediating molecular signals within the cell. Some forms of colon cancer arise when APC is inactivated, though much remains unknown at present about the precise nature of APC inactivation and how this inactivation fosters cancer growth. I was fortunate to participate in a recent study of APC inactivation lead by Drs. J. Amos-Landgraf and W.F. Dove (Amos-Landgraf *et al.* 2012). These investigators used the *Pirc* rat model of human colon cancer in which genetic manipulations were possible, and they obtained a variety of data on the molecular state of colonic tumors that formed in these rats. The analysis I discuss below involved measurements of RNA and DNA allelic ratios at the APC locus from 96 colonic epithelial tissue samples. The samples comprised 27 normal tissues and 69 tumors (Figure 1). Measurements required careful isolation of nucleic acids and advanced pyrosequencing technology to estimate allelic ratios. An important aspect of the experimental protocol was having paired measurements (i.e., both DNA and RNA allelic ratios) on each tissue sample. Of scientific interest was to use these data to learn more about how the cancer cell inactivates the APC gene.

[Figure 1 about here]

The process of developing a statistical model for any data set becomes overly cumbersome if convenient mathematical notation is not provided. So I indexed the $n = 96$ measured tissue samples by a set $\{i\}$ and denoted data $\{(X_i, Y_i)\}$, where

$$\begin{aligned} X_i &= \text{DNA allelic ratio, tissue } i \\ Y_i &= \text{RNA allelic ratio, tissue } i \;. \end{aligned}$$

I viewed these data as random variables, allowing that the particular numbers stored in my computer are realizations of these random variables. Considering the measurement protocol, it is reasonable to assume that (X_i, Y_i) are mutually independent over samples $\{i\}$, and further that they have some common probability distribution reflecting allelic ratio properties over the population of tissues being sampled. The common distribution assumption is problematic because some tissues are normal epithelia and some are from tumors. I clarify the issue by considering some more of the biology in question, and by being more specific about the aims of the statistical analysis.

The biology under investigation concerns modes of inactivation of APC. It is recognized that three distinct processes might be operational over the population of tumors, and so this fact will be reflected in the constructed probability model. Recall first a background point about *Pirc* rats. They carry one defective copy and one normal (wildtype) copy of APC DNA. If the epithelial cells eliminate the wildtype DNA, they gain an advantage during the evolutionary process of tumor formation. One path to tumor formation involves loss of heterozygosity of DNA, which we code as LOH-DNA. Since wildtype RNA can only be transcribed from wildtype DNA, these cells must also have a loss of heterozygosity of RNA, LOH-RNA. In other words the *true* allelic ratios in tumor cells following this pathway are both zero. An alternative pathway to tumor formation would involve inactivation of wildtype RNA by some mechanism other than DNA loss. The cells might *maintain* heterozygosity of DNA (MOH-DNA) but still lose heterozygosity in RNA by some epigenetic silencing (LOH-RNA). Yet a third pathway could entail maintenance of both wildtype DNA and RNA and a disruption during translation of APC. The three natural tumor classes are thus:

class	DNA	RNA	pathway name
1	MOH-DNA	MOH-RNA	translational silencing
2	LOH-DNA	LOH-RNA	DNA loss
3	MOH-DNA	LOH-RNA	epigenetic silencing

An important goal of the experiment was to infer something about the proportion of tumors in the population that follow any of these three developmental pathways. In the absence of measurement error, one could tell immediately from measured allelic ratios what a sample's true allelic ratios were, and hence what its true pathway was. Of course real data come with measurement error, and so a statistical analysis was called for. To proceed, I imagined that each tissue sample is associated with a third random variable, denoted U_i, indicating the pathway taken by tissue i, so $U_i \in \{1, 2, 3\}$. For tissues from the tumor population I assumed there are three probabilities (p_1, p_2, p_3), with

$$P(U_i = j) = p_j \qquad j = 1, 2, 3.$$

The normal tissue samples provide experimental controls, and for these $\{i\}$ I supposed $P(U_i = 1) = 1$, since normal tissue should retain heterozygosity in both DNA and RNA. The probability model becomes fully specified if one further characterizes conditional probability densities $f_j(x, y) = p(x, y|U_i = j)$ for allowable realizations x and y of X_i and Y_i. [Recall that probabilities come from densities by integration, so the probability that the measured DNA allelic ratio is in the set A and the RNA allelic ratio is in the set B, given pathway j, is $\int_A \int_B f_j(x, y)\, dydx$.] This constructs a discrete mixture model for the allelic-ratio data; for tumor tissue i, the associated joint probability density is

$$f(x, y) = \sum_{j=1}^{3} p_j f_j(x, y). \qquad (1)$$

Experienced analysts recognize that competing factors always affect model choice (in this case, the f_js). One seeks models with flexibility to match dominant patterns in the data; but flexibility usually comes with increased numbers of parameters. Large sample sizes may be required to get precise parameter estimates when the parameter space has high dimension. One also seeks models that are plausible by some analysis of the sources of variation. And finally, the demands of statistical inference mean that different models lead to different computational problems. Usually model choice is considered tentative, resulting from an iterative process of specification, inference computation, and diagnostic checking. In the allelic-ratio problem, a choice for f_j that is both convenient and supported by a consideration of the measurement proces is the bivariate normal model. For any real x and y,

$$f_j(x,y) \propto \exp\left\{ -\left(\frac{1}{2}\right) \left[\frac{(x-a_j)^2}{\sigma_{1,j}^2} + \frac{(y-b_j)^2}{\sigma_{2,j}^2} - 2\rho_j \left(\frac{x-a_j}{\sigma_{1,j}}\right)\left(\frac{y-b_j}{\sigma_{2,j}}\right) \right] \right\} \tag{2}$$

This is parameterized by means a_j and b_j, variances $\sigma_{1,j}^2$ and $\sigma_{2,j}^2$, and correlations ρ_j, for $j = 1, 2, 3$. Discrete mixtures of normal distributions are used frequently in applied statistics (e.g., McLachlan and Peel, 2000). Level sets of the component densities are ellipses, as inspection of equation 2 shows, and these are illustrated in Figure 2 when the parameters are fixed at values estimated from the data by the method of maximum likelihood. That is, parameter settings are found to maximize the probability of what has been observed, or equivalently its logarithm l:

$$l = \sum_{i \in \text{tumors}} \log f(x_i, y_i) + \sum_{i \in \text{normals}} \log f_1(x_i, y_i)$$

where the data (X_i, Y_i) are fixed at their observed values (x_i, y_i) and where the marginal density $f(x,y)$ is as in equation (1). Owing to the sum inside the logarithm of f, this estimation problem presents a somewhat complicated computing problem, which I visit next before considering inference summaries derived from the fitted model.

[Figure 2 about here]

Discrete mixture models form the quintessential application of the Expectation Maximization (EM) algorithm, perhaps the most important statistically-based optimization method. Briefly, EM proceeds by first considering the logarithm of the likehood function one would compute if one were lucky enough to also observe the $\{U_i\}$. A mathematical form of the *complete-data* log-likelihood l_c is readily available, but the function itself cannot be evaluated because in tumor tissues one lacks the missing U_i:

$$l_c = \sum_i \sum_j 1[U_i = j] \left[\log p_j + \log f_j(x_i, y_i)\right].$$

EM proceeds by maximizing instead the conditional expectation of l_c given the available data, using a current guess at the parameter values needed to compute the conditional expectation. This general process is described by McLachlan and Peel (2000) and is elaborated by Fraley and Raftery (2002) in the context of Gaussian mixture models. Indeed, Fraley and Raftery's `mclust` software system provides a powerful and user-friendly approach to mixture-model analysis. For the allelic-ratio example I would have used `mclust` had it not been for further intrinsic constraints on the parameter values suggested by the genetic context.

Considering the effects of LOH or MOH on allelic ratios, one expects, for example, that the DNA content of cells that are MOH-DNA ought to have the same average value, regardless of whether the cell's path to tumor formation follows the MOH-DNA/MOH-RNA pathway or the MOH-DNA/LOH-RNA pathway. This imposes constraints on model parameters, since different classes now share aspects of the mean vector of their bivariate distribution:

class	pathway		mean DNA	mean RNA
1	MOH-DNA	MOH-RNA	a_1	b_1
2	LOH-DNA	LOH-RNA	a_2	b_2
3	MOH-DNA	LOH-RNA	a_1	b_2

I imposed no special constraints on the variance/correlation parameters nor on the mixing probabilities (p_1, p_2, p_3), and I deployed an EM algorithm to derive maximum likelihood estimates (MLEs). Ellipses in Figure 2 express the MLEs of class means and covariance structures; for example class 2 exhibits a positive correlation between X_i and Y_i owing to the positive slope of the regression line in that class.

Parameter estimates not indicated by Figure 2 are the estimated mixing proportions, which I computed to be: $(\hat{p}_1, \hat{p}_2, \hat{p}_3) = (0.143, 0.713, 0.144)$. These provide the first main inference summary. I estimated for the population of tumors under study that the three developmental pathways are traversed in these relative proportions; prior to the Amos-Landgraf report it was known that the three named pathways *could* explain tumor development, but a quantitative assessment of prevalence had not been available. Notably, most tumors (71.3%) are estimated to have lost APC function via loss of the wildtype APC DNA, as opposed to some alternative mechanism. And 14.4% are estimated to have undergone an epigenetic silencing pathway that inactivates the wildtype APC RNA. These estimates and their interpretation enhanced the contribution of the Amos-Landgraf report.

In addition to estimates of pathway prevalence for the entire tumor population, the mixture model analysis produces tumor-specific inferences about which pathways were probably traversed. Statistically, this yields a clustering of the tumors into three groups. The contour lines in Figure 1 summarize tumor-specific posterior probabilities for the epigenetic silencing pathway:

$$\begin{aligned} \text{post}(x, y) &= P(U_i = 3 | X_i = x, Y_i = y) \\ &= \frac{p_3 f_3(x, y)}{f(x, y)} \end{aligned}$$

where objects are as previously defined, and where Bayes's rule converts the forward sampling probabilities into posterior probabilities necessary for inference. Lines in Figure 1, drawn at various reference levels of this posterior probability, show which tumors we estimate to have followed the epigenetic silincing pathway. A second model-based analysis was used by Amos-Landgraf *et al.* to further assess within-tumor heterogeneity; I encourage the reader to examine the supplementary material of that paper in order to appreciate the more refined inferences that are possible within the mixture-model framework. Ultimately, the model-based computations provide a useful summary of experimental data and a guide forward towards follow-up experiments. Because they ignore detailed mechanistic elements of the biology, there is no way that the simple mixture models discussed above provide a comprehensive, error-free assessment of the system under study, and this limitation must always be considered lest one risks over-interpreting the numerical findings. Model diagnostics and a fluid interaction among investigators helps to mitigate the effects of oversimplification.

3 Karyotypic abnormalities in stem cells

Owing to their great potential to advance medical therapies, stem cells have been the focus of intense research. Both human embryonic stem cells (ESCs) and induced pluripotent cells (iPSCs) have self-renewal capacity as well as the capacity to differentiate into any of the body's cell lineages. Investigators who use stem-cell stocks routinely have their cells tested for the presence of karyotypic abnormalities, which represent large-scale genomic changes that may have arisen during growth in culture. For example, having three copies of chromosome 12 (trisomy 12), is a relatively frequent stem-cell abnormality. Dr. Karen Montgomery's lab in Madison, WI, has been a central stem-cell testing facility for this purpose, and from her lab, Taapken *et al.* (2011) reported karyotype data from 1715 stem-cell cultures, these being comprised of 1163 ESC cultures and 552 iPSC cultures tested in their facility. In addition to providing useful data to stem-cell researchers, Taapken *et al.* addressed a critical question: Do ESCs and iPSCs differ significantly in their propensity to accumulate genomic abnormalities? Ever since iPSCs were discovered, there has been debate about their utility for biomedical research, and so there was great interest in the study findings.

I was invited to participate in the karyotype data analysis, which seemed somewhat elementary at first. From data in Table 1 of Taapken *et al.*, there are $m = 1163$ ESC cultures, of which $x = 150$ show some kind of abnormal karyotype; similarly there are $n = 552$ iPSC cultures of which $y = 69$ are abnormal. Evidently, the rates $x/m = 12.9\%$ and $y/n = 12.5\%$ are quite close. To be more precise, x and y are realizations of random counts X and Y. One might reason that X and Y have binomial distributions, since they represent sums of Bernoulli trials that are reasonably thought to be independent, based on the nature of the karyotype assay, and then compute a $p-$value. Recall, the $p-$value measures the probability of something as or more extreme a difference than the observed difference, in hypothetical repeats of the study, and assuming no real difference in underlying rates. Of course, in comparing binomial counts, either Fisher's exact test or Pearson's chi-square test would do (e.g. Agresti, 1990, page 59), and both indicate that 12.9% and 12.5% are not *significantly* different ($p-$value $= 0.9$).

I might not be discussing the case if that was the end of the story! Looking deeper, there were several important issues not accounted for in the above analysis. On one hand, there was a variety of abnormalities and investigators wanted to look specifically at the components comprising the overall aberration rate. More substantially, a potentially important factor called *passage* was not accounted for, though it could easily have affected the comparison. Roughly speaking, passage refers to the age of the cells in culture. As cells grow *in vitro*, they naturally reach the limits of their containing vessels; cells are passaged when a small sample is transferred to a new vessel. Most of the cell cultures considered in Taapken *et al.* included data on the passage number at the point when the karyotype was measured (1662 of 1715 cultures). Figure 3 summarizes the passage data for the $m' = 1128$ ESC cultures and the $n' = 534$ iPSC cultures for which data were available. Evidently, the iPSC cultures were significantly younger (i.e., measured at significantly lower passage number) than the ESC cultures. Not accounting for passage created a problem because aberration rate ought naturally to increase with passage, if it is related at all; in reality, it might be that iPSC cells have a significantly higher rate of abnormality, but they appear similar to ESC cells because of the imbalance in the timing of the measurements. Fortunately, the available data structure permitted a more refined statistical analysis, since on most cultures both aberration and passage information were measured.

[**Figure 3 about here**]

It is helpful to introduce some notation before proceeding further. Consider the set $\{i\}$ indexing the cultures for which both aberration data and passage data are available. Suppose also in most of what follows that a specific aberration (e.g., trisomy 12) is in focus, which is to be compared between ESC and iPSC. Let Y_i be the Bernoulli trial indicating whether ($Y_i = 1$) or not ($Y_i = 0$) culture i is observed to have the aberration in question. Let Z_i be the passage variable, recording the passage number at the time the karyotype is measured, and let X_i denote the cell type (ESC or iPSC). Thus each culture i provides data (X_i, Y_i, Z_i). The analysis deployed in Taapken *et al* and reviewed here proceeded using a model for the conditional probability

$$\pi(x, z) = P(Y_i = 1 | X_i = x, Z_i = z), \tag{3}$$

for both cell types x and all passages z. Considering the relatively large sample size (1662), I suspected that a parametric model might be too restrictive and further that there might be sufficient information to reliably use something *nonparametric*. As an aside, an obvious parametric model would be logistic regression, in which $\log\{\pi(x,z)/[1-\pi(x,z)]\} = a+bx+cz$ for parameters a, b, c. Instead, I developed a model by reasoning as is often done in event-time modeling. I supposed that each culture i is associated with a latent *true* event time T_i that marks when the cells first incurred the named karyotypic aberration, considered relative to a lifespan of cells from initial establishment to some time well beyond the present, and measured on the passage scale. It is generally understood that cells start life in some *normal* state; any incurred damage is irreversible. Thus, the data structure is that of *current status* or *type-I interval censored* event-time data (e.g., Huang and Wellner, 1997). Specifically, observing $Y_i = 1$ is equivalent to knowing $T_i \leq Z_i$; the onset time of the aberration must have preceded the passage time-stamp on the cells when their karyotype was measured, else $Y_i = 0$. The true onset times $\{T_i\}$ are unobserved, but their cumulative distribution functions (c.d.f.'s), at least conditionally upon cell type X_i, relate to the aberration probabilities in (3), since

$$\begin{aligned} \pi(x, z) &= P(T_i \leq z | X_i = x, Z_i = z) \\ &= F_x(z), \end{aligned} \tag{4}$$

where $F_x(z)$ is the c.d.f. of T_i for cells of type $X_i = x$. The statistics student will recall the central importance of the c.d.f. in characterizing all probability statements about a random variable [i.e., the c.d.f. is a non-decreasing function with range $[0,1]$ that at each point in its domain is continuous from the right, and that has limits from the left.] I have also made the seemingly innocuous assumption that Z_i and T_i are conditionally independent given X_i. The scientific question regarding differences in aberration rate between ESC and iPSC amounts to testing a hypothesis about the c.d.f.'s F_{E} and F_{iP}, these both being functions of passage z. Specifically, the null hypothesis under test is

$$H_0 : F_{\mathrm{E}}(z) = F_{\mathrm{iP}}(z) \quad \text{for all passages } z \; . \tag{5}$$

On H_0, and at any passage z, the probability that a cell culture acquires the damage in question does not depend on the cell type. Deploying a test of H_0 requires a test statistic and an estimate of this statistic's null distribution. I developed a likelihood-based statistic in Taapken *et al.*, even though the c.d.f.s were not restricted to a parametric form, and I used bootstrap sampling to compute p-values. As a function of two c.d.f's F_{E} and F_{iP}, consider the log likelihood

$$\begin{aligned} l\left(F_{\mathrm{E}}, F_{\mathrm{iP}}\right) &= \log\left[\prod_i P\left(Y_i = y_i | X_i = x_i, Z_i = z_i\right)\right] \qquad (6) \\ &= \sum_{x \in \{\mathrm{E,iP}\}} \sum_{i: x_i = x} \left\{ y_i \log F_x(z_i) + (1 - y_i) \log\left[1 - F_x(z_i)\right]\right\} \\ &= l_{\mathrm{E}}\left(F_{\mathrm{E}}\right) + l_{\mathrm{iP}}\left(F_{\mathrm{iP}}\right). \end{aligned}$$

To parse this log likelihood, recall that independent observations have their probabilites entering by multiplication, and hence additively on the log scale. Further, the sum separates into two parts based on the source cell type X_i of culture i. Within cell type, the sum captures the log of a Bernoulli likelihood in which the response probability fluctuates according to passage through the respective c.d.f.. It is a classic statistics problem, whose solution goes back to Ayer *et al.* (1955), to maximize either component l_{E} or l_{iP} over the class $\mathcal{F}$ of all possible c.d.f.'s over passage. Indeed, the *pool adjacent violators algorithm* proposed by Ayer *et al.* delivers estimates

$$\hat{F}_{\mathrm{E}} = \arg\max_{F \in \mathcal{F}} l_{\mathrm{E}}(F) \quad \text{and} \qquad \hat{F}_{\mathrm{iP}} = \arg\max_{F \in \mathcal{F}} l_{\mathrm{iP}}(F)$$

Figure 4 shows these two nonparametric maximum likelihood estimates from the Taapken *et al.* data, when considering Y_i to be the indicator of any form of karyotypic abnormality.

[Figure 4 about here]

With sights focused on the inference goal, one needs to construct some test statistic to evaluate (5); that is to construct a statistic that measures a difference between ESC and iPSC aberration rates while accounting for differences in passage distribution between the cell types. I used the likelihood ratio statistic

$$T = l\left(\hat{F}_{\mathrm{E}}, \hat{F}_{\mathrm{iP}}\right) - l\left(\hat{F}_0, \hat{F}_0\right)$$

where l is the log-likelihood in (6), where $\hat{F}_{\mathrm{E}}$ and $\hat{F}_{\mathrm{iP}}$ are maximum likelihood estimates in the unconstrained model, and where $\hat{F}_0$ is the maximum likelihood estimate of the common distribution on H_0. Naturally, $\hat{F}_0$ is calculated by pooling the ESC and iPSC data and appying Ayer's algorithm. Numerically, I obtained the value $t = 5.35$ as the realization of T in Taapken's data. Measuring the extent to which the event $T = 5.35$ provides evidence against H_0 is the final step of the hypothesis testing exercise. Indeed I know of no mathematical result that would characterize the probability distribution of T, however this distribution can also be estimated numerically using bootstrap sampling (e.g., Davison and Hinkley, 1997). One treats the estimated null distribution $\hat{F}_0$ as if it were true, and one simulates hypothetical new data sets from this fitted model. The statistic is recomputed on every simulated data set, providing a Monte Carlo approximation to the estimated distribution of T. In the case under study, I found that 23% of null simulated data sets lead to a statistic as or more extreme that $t = 5.35$ (Figure 5). As this p-value is quite moderate, one finds no evidence against the null hypothesis that ESC and iPSC cells incur aberrations at the same rate.

[**Figure 5 about here**]

Coda

Science is like climbing
Statistics is like rope
Often, it's for safety
Sometimes, you can't get there without it

Acknowledgements

The author thanks Christina Kendziorski for help in the development of the work presented here.

References

1. Agresti, A (1990). *Categorical data analysis.* John Wiley & Sons. New York.

2. Amos-Landgraf, JM, Irving, AA, Hartman, C, Hunter, A, Laube, B, Chen, X, Clipson, L, Newton, MA, and Dove, WF (2012). Monoallelic silencing and haploinsufficiency in early murine neoplasms. *Proc. Natl. Acad. Sci.*, 109, 2060-2065.

3. Ayer, M, Brunk, HD, Ewing, GM, Reid, WT, and Silverman, E (1955). An empirical distribution function for sampling with incomplete information. *Ann. Math. Statist.*, 26, 641-647.

4. Davison, AC and Hinkley, DV (1997). *Bootstrap methods and their application.* Cambridge University Press, Cambridge, UK.

5. Fraley, C and Raftery AE (2002). Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association*, 97, 611-631.

6. Huang, J and Wellner, JA (1997). Interval censored survival data: A review of recent progress. *Proc 1st Seattle Symp in Biostatistics: Survival Analysis.* Eds. D. Lin and T. Fleming. Springer-Verlag, New York.

7. McLachlan, G and Peel, D (2000). *Finite Mixture Models*, John Wiley & Sons. New York.

8. Taapken, SM, Nisler, BS, Newton, MA, Sampsell-Barron, TL, Leonhard, KA, McIntire, EM, and Montgomery, KD (2011). Karyotypic abnormalities in human induced pluripotent stem cells and embryonic stem cells. *Nature Biotechnology*, 29, 313-314.

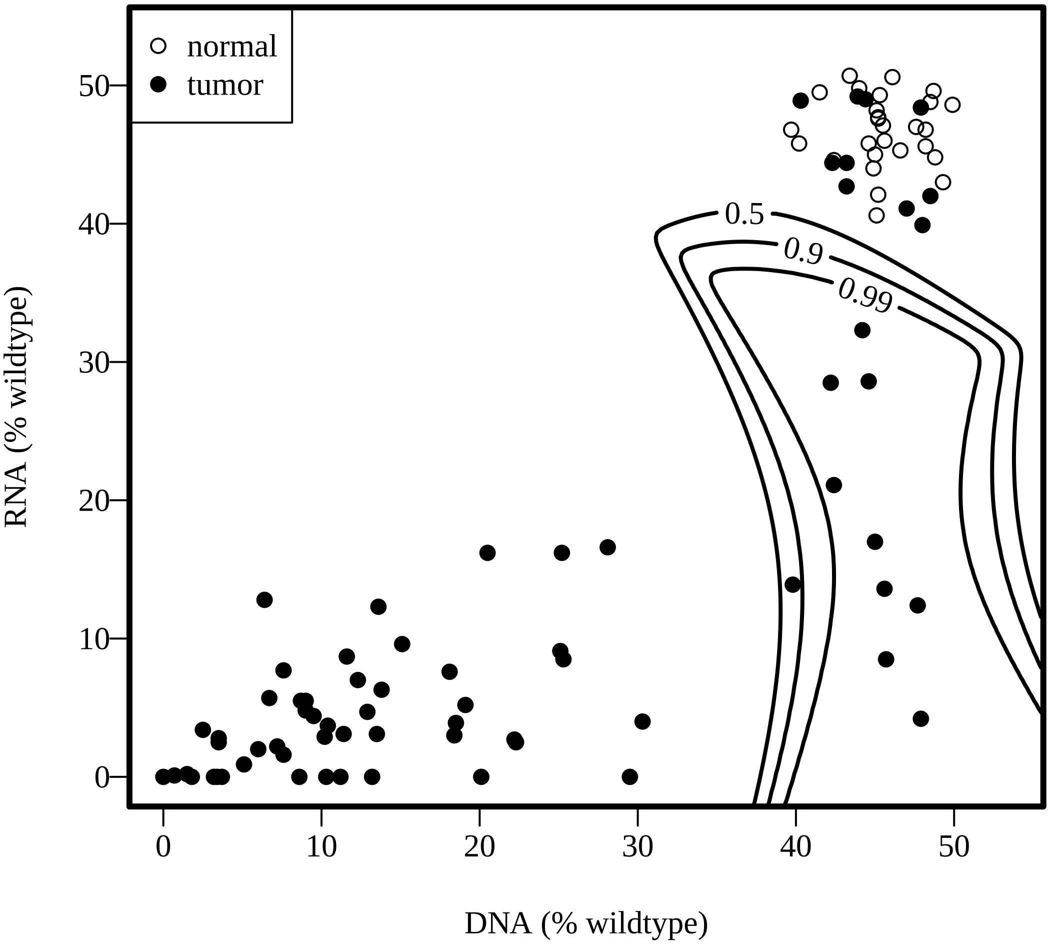

Figure 1: Allelic ratios (percentage of the wild-type allele) from 96 colonic tissue samples in *Pirc* rats, in both DNA and RNA, as in Amos-Landgraf *et al.* 2012. Contour lines are derived from a statistical analysis of these data and reflect properties of an estimated three-class mixture model.

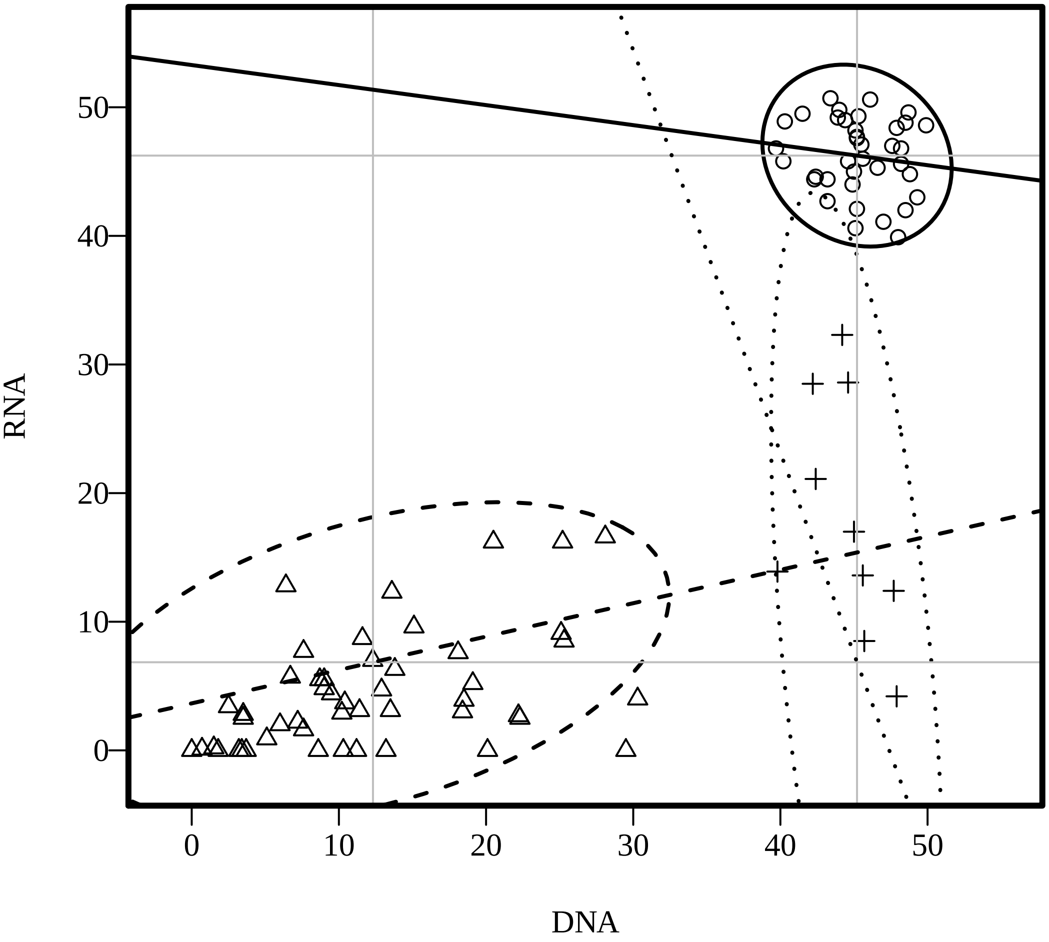

Figure 2: Components of the mixture model fitted to the Amos-Landgraf data. Each of the three bivariate Gaussian component distributions has a centering point (at one of the intersections of the grey lines), an ellipse centered at that point marking a level set of the joint density function, and a regression line. Each ellipse marks 95% of the probability content of the component distribution, and regression lines indicate the expected RNA ratio given the component and the DNA ratio. Note that prior to fitting the three components were allowed to have arbitrary covariance matrices, but they were constrained to have means expressing elements of the problem structure. The original data points are also plotted and marked according to the component that they are most likely to have been drawn from.

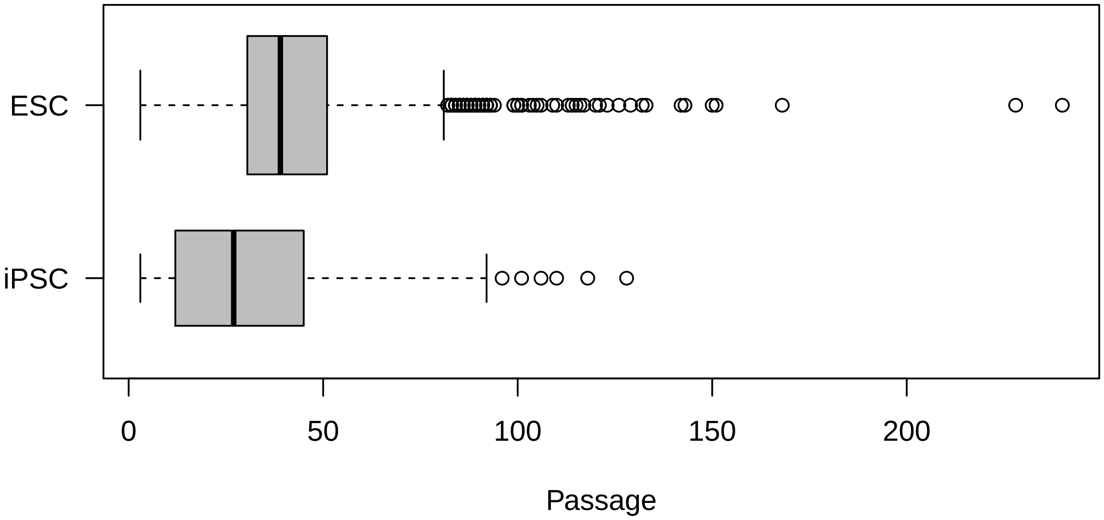

Figure 3: Boxplots compare the empirical distribution of passage data between embryonic stem cell cultures and induced pluripotent stem cell cultures. Recall that in a boxplot, each box contains the middle 50% of the data values, with the vertical bar at the median, and with whiskers extending out to 1.5 times the inter-quartile range, if data points lie beyond that amount, or to the most extreme observation, otherwise. ESC and iPSC cell cultures differ in their distribution of passage number.

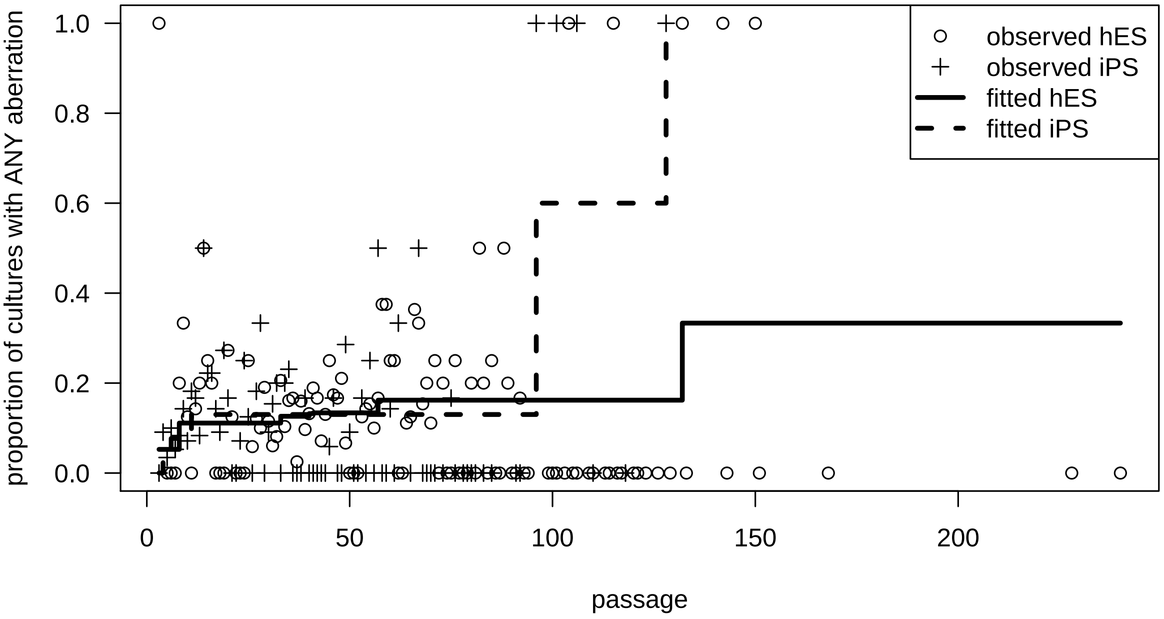

Figure 4: Nonparametric maximum likelihood estimates $\hat{F}_{\mathrm{E}}$ and $\hat{F}_{\mathrm{iP}}$, the distributions of time to aberration, when estimating the rates of *any* karyotypic abnormality in stem cell cultures. Data are also shown in a summary format in which at each unique passage value we record the proportion of abnormal cultures among all cultures observed at that passage.

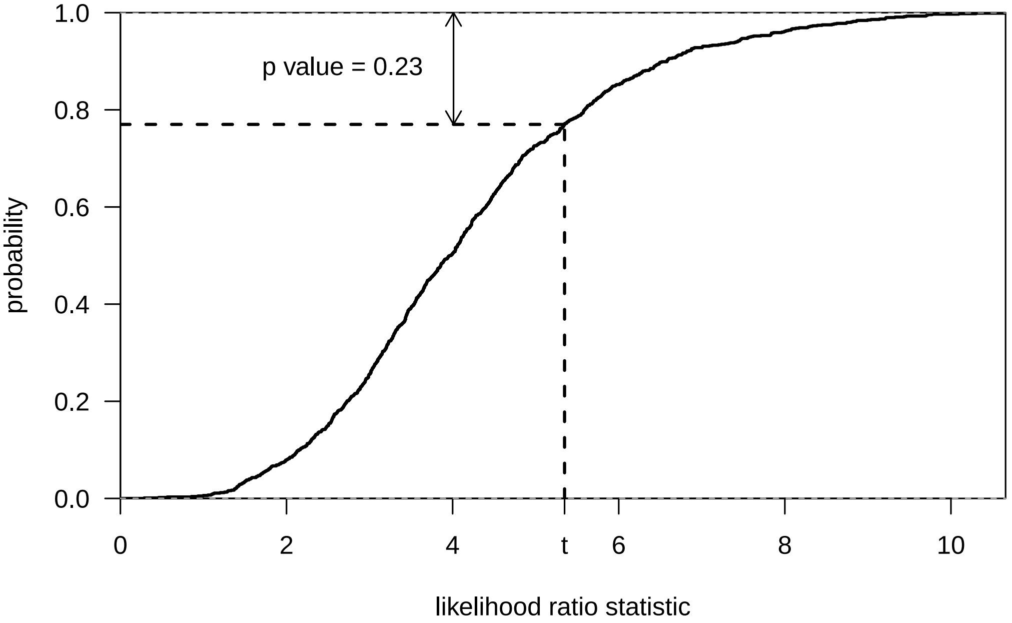

Figure 5: Bootstrap distribution of the likelihood ratio statistic in the stem cell case study. The null hypothesis on test asserts that the distribution for time to aberration is common to both ESC and iPSC. This common distribution was estimated by nonparametric maximum likelihood via Ayer's algorithm applied to the combined data set, and this estimated distribution $\hat{F}_0$ was the basis for a bootstrap simulation. Repeatedly, in $B = 1000$ trials, two (ESC and iPSC) bootstrap data sets were generated by drawing independent Bernoulli trials, using fixed passage data and using $\hat{F}_0$ with these passages to determine the success probability of the trials. A log-likelilood ratio statistic was computed from the ESC and iPSC bootstrap data sets so generated, using Ayer's algorithm separately in ESC and iPSC conditions to obtain a statistic. The empirical distribution of these 1000 statistics is shown above; lines indicate the position of the statistic calculated from Taapken *et al.* data ($t = 5.35$), and thus the resulting p-value equals 0.23.