Some Elementary Statistical Inferences
(Hogg Chapter Four)

STAT 405-01: Mathematical Statistics I *

Fall Semester 2015

Contents

Tuesday 3 November 2015
– Read Section 4.1 of Hogg

1 Samples, Statistics, and Estimators

1.1 (Frequentist) Statistical Inference

So far, our consideration of probabilities and random variables has involved taking a known distribution and using it to calculate the probabilities of various equations and inequalities for random variables. Now we're going to consider the case where

we have some random variables X_1, X_2, etc, for which the probability distribution is at least partially unknown, and developing a prescription that allows us to make some statements about the unknown quantities, given a particular set of values $X_1 = x_1$, $X_2 = x_2$, etc.

One possibility is that the probability distribution has some unknown parameters, e.g., we have a Poisson distribution with unknown mean, or Gamma distribution with unknown α and β. If we call these parameters θ_1, θ_2, etc., and we combine the observable random variables into a vector $\mathbf{X}$ and the parameters into another vector $\boldsymbol{\theta}$, we have a probability distribution

$$f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta}) \tag{1.1}$$

(We've written this as a pdf for a continuous random vector, but the formalism is similar if some or all of the random variables are discrete.) This gives us the probability density at $\mathbf{X} = \mathbf{x}$, for specific values of the parameters $\boldsymbol{\theta}$. The idea is that when we perform the experiment, we get specific values $\mathbf{x}$, and we can then use $f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta})$ to say something the value(s) of $\boldsymbol{\theta}$.

A case of particular interest is where the $\{X_i\}$ are independent and identically distributed, i.e., they form a *random sample*. Then

$$f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta}) = \prod f(x_i;\boldsymbol{\theta}) \tag{1.2}$$

1.2 The Bayesian Point of View

The classical perspective, known as the frequentist interpretation, that we're using here, is a bit odd. The *sampling function* $f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta})$ is not a distribution for values of $\boldsymbol{\theta}$; it's a distribution for $\mathbf{x}$ values, which happens to depend on the values of $\boldsymbol{\theta}$. This makes it less than straightforward to infer something about $\boldsymbol{\theta}$ from the form of $f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta})$ evaluated at the actual observed $\mathbf{x}$, as a function of $\boldsymbol{\theta}$.

On the other hand, the Bayesian approach treats the inherently random realization $\mathbf{x}$ and the unknown parameters $\boldsymbol{\theta}$ on equal footing; both represent uncertainty or lack of knowledge. The sampling function is thus better written $f(\,\mathbf{x} \mid \boldsymbol{\theta}\,)$, a probability density for $\mathbf{x}$ given parameter values $\boldsymbol{\theta}$. (If you want to think in the language of random variables, you could consider a case where a meta-experimenter sets up the environment by randomly choosing parameters $\boldsymbol{\theta}$, which are the realization of a random vector $\boldsymbol{\Theta}$ with some distribution $f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$; we'll suppress all of these subscripts since we'll know what quantities we're talking about from the arguments.) If your a priori knowledge of the parameters $\boldsymbol{\theta}$ is reflected by a probability distribution $f(\boldsymbol{\theta})$, then we could think about the joint probability density to have parameters $\boldsymbol{\theta}$ and observed data $\mathbf{x}$, which is

$$f(\boldsymbol{\theta},\mathbf{x}) = f(\,\mathbf{x} \mid \boldsymbol{\theta}\,)f(\boldsymbol{\theta}) \tag{1.3}$$

We're interested in the a posteriori knowledge of $\boldsymbol{\theta}$, given that we made the measurements $\mathbf{x}$, which is a considitional probability:

$$f(\,\boldsymbol{\theta} \mid \mathbf{x}\,) = \frac{f(\boldsymbol{\theta},\mathbf{x})}{f(\mathbf{x})} = \frac{f(\,\mathbf{x} \mid \boldsymbol{\theta}\,)f(\boldsymbol{\theta})}{f(\mathbf{x})} \tag{1.4}$$

This is a form of Bayes's theorem. The parts of it have names:

- $f(\boldsymbol{\theta})$ is the prior pdf for the parameters
- $f(\,\boldsymbol{\theta} \mid \mathbf{x}\,)$ is the psoterior pdf for the parameters
- $f(\,\mathbf{x} \mid \boldsymbol{\theta}\,)$ is the sampling function, which is also called the likelihood function when we consider its $\boldsymbol{\theta}$ dependence.
- The denominator $f(\mathbf{x}) = \int f(\,\mathbf{x} \mid \boldsymbol{\theta}\,)f(\boldsymbol{\theta})\,d\boldsymbol{\theta}$ is a normalization factor.

There are two challenges which keep everyone from using the Bayesian approach: first, there is the conceptual subtlety of assigning probabilities to things which are only unknown and

not the result of repeatable experiments. Second, we have to make some determination of the starting point, the prior pdf $f(\boldsymbol{\theta})$, which can be somewhat arbitrary.

1.3 Statistics and Estimators

Given a random sample (or any random vector) $\mathbf{X}$, a statistic $T(\mathbf{X})$ is any function of the random variables $\{X_i\}$. The statistic $T(\mathbf{X})$ is itself a random variable; given a particular realization (set of data/values) $\mathbf{x}$ of the random vector $\mathbf{X}$, the statistic takes on a numerical value $T(\mathbf{x})$. Often, a statistic can be used to get an estimate of one of the parameters. We say that $T(\mathbf{X})$ is an estimator for the parameter θ.

1.3.1 Unbiased (and Biased) Estimators

A statistic $T(\mathbf{X})$ is called an unbiased estimator of the parameter θ if $E[T(\mathbf{X})] = \theta$. For example, we know that if $\mathbf{X}$ is a sample of size n drawn from any distribution, the statistic $\overline{X} = \sum_{i=1}^{n} X_i$ satisfies $E(\overline{X}) = \mu$ where $E(X_i) = \mu$. So for example, given a sample drawn from a $N(\mu, \sigma^2)$ normal distribution, $\overline{X}$ is an unbiased estimator of μ.

You may wonder why we'd ever use a biased estimator, but bias is more insidious than you'd think. For example, consider an exponential distribution with unknown parameter λ:

$$f(x;\lambda) = \lambda e^{-\lambda x} \qquad 0 < x < \infty \tag{1.5}$$

We know that $\frac{1}{\lambda}$ is the expectation value of an exponential random variable, so $\overline{X}$ is an unbiased estimator of $\frac{1}{\lambda}$. However, one can show that $\frac{1}{\overline{X}}$ is a biased estimator of λ, because in general

$$E\left(\frac{1}{\overline{X}}\right) \neq \frac{1}{E(\overline{X})} \tag{1.6}$$

In fact, we can calculate this expectation value for a sample of size n drawn from an exponential distribution with parameter λ:

$$\begin{aligned} E\left(\frac{1}{\overline{X}}\right) &= \int_0^\infty \cdots \int_0^\infty \frac{n}{x_1 + \cdots + x_n} \lambda e^{-\lambda x_1} \cdots \lambda e^{-\lambda x_n}\, dx_1 \cdots dx_n \\ &= n\lambda^n \int_0^\infty \cdots \int_0^\infty \frac{1}{x_1 + \cdots + x_n} e^{-\lambda(x_1 + \cdots + x_n)}\, dx_1 \cdots dx_n \end{aligned} \tag{1.7}$$

It's possible, by changing variables from $\{x_i\}$ to $\{y_i\}$ where

$$y_i = \lambda \sum_{j=1}^{i} x_j \tag{1.8}$$

so that the limits of the integrals are

$$0 < y_i < y_{i+1},\ i < n; \qquad 0 < y_n < \infty \tag{1.9}$$

to show that

$$E\left(\frac{1}{\overline{X}}\right) = n\lambda \int_0^\infty \frac{e^{-y_n}}{y_n} \frac{(y_n)^{n-1}}{\Gamma(n)}\, dy_n = n\lambda \frac{\Gamma(n-1)}{\Gamma(n)} = \frac{n}{n-1}\lambda \tag{1.10}$$

Of course, in this case we can just multiply by the constant vector to get

$$\left(1 - \frac{1}{n}\right) \frac{1}{\overline{X}} \tag{1.11}$$

as an unbiased estimator for λ.

1.3.2 Maximum Likelihood

One method for obtaining an estimator for a parameter θ is to pick the value which maximizes the likelihood $L(\theta) = f(\mathbf{x};\theta)$ as a function of θ. This seems reasonable, since it gives the

parameter value (or values in the case where $\boldsymbol{\theta}$ is a collection of parameters) for which the actual data were most likely. But there is still something a bit fishy, since the likelihood function is not a probability distribution for θ; it's a probability distribution for $\mathbf{x}$ whose form depends on the parameter θ. Still, we refer to the value of θ which maximizes $f(\mathbf{x};\theta)$ as $\widehat{\theta}$. If the parameter is continuous, this is defined by a derivative

$$\frac{d}{d\theta} f(\mathbf{x};\theta)\bigg|_{\theta=\widehat{\theta}} = 0 \tag{1.12}$$

The Bayesian picture is a little more logical, and gives a sense of why maximum likelihood is a good idea. In the Bayesian case, we end up with a posterior probability distribution $f(\,\theta \mid \mathbf{x}\,)$ for θ based on the data $\mathbf{x}$ which we observed. The maximum of the posterior occurs at the value of θ which is the mode of this distribution. Since Bayes's theorem tells us

$$f(\,\theta \mid \mathbf{x}\,) \propto f(\,\mathbf{x} \mid \theta\,) f(\theta) \tag{1.13}$$

we see that if the prior probability density for θ is uniform, the maximum likelihood value $\widehat{\theta}$ is the value which maximizes the posterior $f(\,\theta \mid \mathbf{x}\,)$.

The value of the maximum likelihood estimate $\widehat{\theta}$ depends on $\mathbf{x}$, so we can also think of a maximum likelihood estimator $\widehat{\theta}$ which is a random variable, in fact a statistic constructed from the random sample $\mathbf{X}$.

In practice, it is easier to calculate the maximum likelihood estimate using the logarithm $\ell(\theta) = \ln L(\theta)$ of the likelihood function. As an example, consider once again the exponential distribution. The likelihood function is

$$L(\lambda) = f(\mathbf{x};\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} \tag{1.14}$$

so the log-likelihood is

$$\ell(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i \tag{1.15}$$

and its derivative is

$$\ell'(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i \tag{1.16}$$

setting this to zero gives the estimator

$$\widehat{\lambda} = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\overline{X}} \tag{1.17}$$

which we've noted is biased.

Thursday 5 September 2015 – Review for Prelim Exam Two

The exam covers materials from weeks 5 and 7-10 of the course, i.e., Hogg sections 2.6-2.7 and 3.1-3.6, and problem sets 5-8.

Tuesday 10 November 2015 – Second Prelim Exam

Thursday 12 November 2015 – Read Section 4.2 of Hogg

2 Interval Estimation

2.1 Confidence Intervals

Estimators of unknown parameters provide us with a method to get a single number given an instance of a random sample.

This is also known as point estimation; it's also useful, however, to generate an interval which is designed to contain the true value for some fraction of sample realizations. This is known as a (frequentist) confidence interval. It's a pair of statistics $L = L(\mathbf{X})$ and $U = U(\mathbf{X})$ chosen so that the probability that the parameter θ lies between them is $1 - \alpha$ (e.g., if $\alpha = 0.10$, it is 90%):[1]

$$P(L < \theta < U) = 1 - \alpha \tag{2.1}$$

It's important to note that the probabilities here refer to the randomness of L and U, and not to the unknown θ. From the frequentist perspective, we can't talk about probabilities for different values of θ; it has some specific value, even if it's unknown. What's random is the sample $\mathbf{X}$ and the statistics L and U created from it.

Given a particular realization $\mathbf{x}$ of the sample $\mathbf{X}$, we have a specific confidence interval between $\ell = L(\mathbf{x})$ and $u = U(\mathbf{x})$. Note that the probabilistic statements do not actually refer to the properties of a particular confidence interval (ℓ, u) but to the procedure used to construction of the confidence interval.

One method to construct the confidence interval is to choose a statistic $T = T(\mathbf{X}; \theta)$, known as a *pivot variable*, whose probability distribution is a known function of the parameters, and construct an interval using the percentiles of the distribution

$$P(a < T(\mathbf{X}; \theta) < b) = 1 - \alpha \tag{2.2}$$

By algebraically solving the inequalities $a < T(\mathbf{X}; \theta)$ and $T(\mathbf{X}; \theta) < b$ for θ, we should be able to write

$$P(L(\mathbf{X}) < \theta < U(\mathbf{X})) = 1 - \alpha \tag{2.3}$$

[1] We're implicitly considering a *two-sided* confidence interval, so we also have $P(\theta < L) = \alpha/2$ and $P(U < \theta) = \alpha/2$.

Note that this construction is not unique; different choices for the pivot variable will give different confidence intervals with the same confidence.

2.2 Aside: Bayesian Plausible Intervals

In the Bayesian perspective, where a particular sample instance $\mathbf{x}$ results in a posterior probability distribution $f(\,\theta \mid \mathbf{x}\,)$, we would have a straightforward and unique definition of a plausible interval (ℓ, u)

$$\int_{\ell(\mathbf{x})}^{u(\mathbf{x})} f(\,\theta \mid \mathbf{x}\,)\, d\theta = 1 - \alpha \tag{2.4}$$

This would be an interval such that the unknown value of θ had a $1 - \alpha$ probability of lying within it.[2] This would be a unique prescription given the posterior pdf, but recall that to get that pdf, we need to use whatever information we have about the problem to construct the appropriate prior pdf $f(\theta)$.

2.3 Example: Mean of a Normal Distribution

To illustrate the pivot variable method, consider the case where $\mathbf{X}$ is a sample of size n drawn from a $N(\mu, \sigma)$ distribution with both μ and σ unknown, where we want a confidence interval on μ. The pivot variable should depend on μ and $\mathbf{X}$ but not σ, so

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \tag{2.5}$$

will not work, even though we know it obeys as $N(0, 1)$ distribution (because $\overline{X}$ obeys a normal distribution with $E(\overline{X}) = \mu$ and

[2] Again, we're assuming it's two-sided so that $P(\,\theta < \ell(\mathbf{x}) \mid \mathbf{X} = \mathbf{x}\,) = \alpha/2$ and $P(\,u(\mathbf{x}) < \theta \mid \mathbf{X} = \mathbf{x}\,) = \alpha/2$.

$\mathrm{Var}(\overline{X}) = \sigma/\sqrt{n}$. Fortunately, we know from Student's theorem that

$$T = \frac{\overline{X} - \mu}{\sqrt{S^2/n}} \tag{2.6}$$

obeys a t distribution with $n-1$ degrees of freedom. This will work as a pivot variable, since

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{2.7}$$

depends only on the sample, and requires no knowledge of μ or σ. Having identified a pivot variable which obeys a t distribution is useful not so much because we know the precise form of the pdf

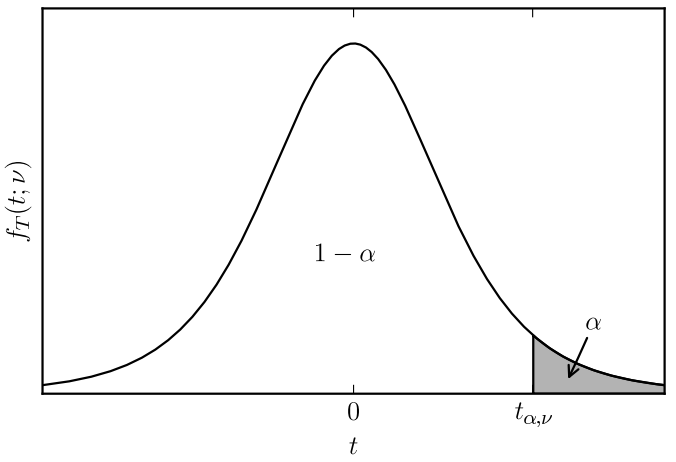
$$f_T(t;\nu) = \frac{\Gamma([\nu+1]/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)}\left(1 + \frac{t^2}{\nu}\right)^{-[\nu+1]/2} \tag{2.8}$$

but because it's a standard distribution for which the percentiles are tabulated in various books or available in R, scipy, etc. The 90th percentile, for example, of a t distribution with ν degrees of freedom is written $t_{0.1,\nu}$; in general, the $(1-\alpha)\times 100$th percentile $t_{\alpha,\nu}$ is defined by

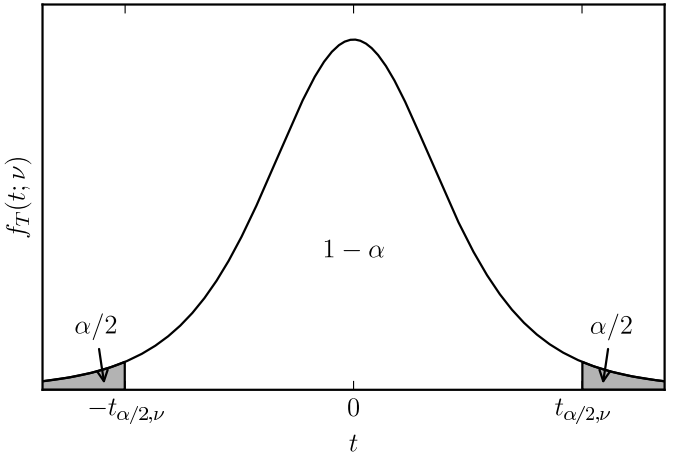
$$1 - \alpha = P(T \le t_{\alpha,\nu}) = \int_{-\infty}^{t_{\alpha,\nu}} f_T(t;\nu)\,dt \tag{2.9}$$

or equivalently by

$$\int_{t_{\alpha,\nu}}^{\infty} f_T(t;\nu)\,dt = \alpha \tag{2.10}$$



Since we want a two-sided confidence interval, we actually need $t_{\alpha/2,\nu}$ and $t_{1-\alpha/2,\nu}$. Since the t distribution is symmetric, though, we can take advantage of the fact that $t_{1-\alpha/2,\nu} = -t_{\alpha/2,\nu}$, e.g., the 5th percentile is minus the 95th:



Thus, returning to the case of the pivot variable T, which is t-distributed with $n-1$ degrees of freedom,

$$\begin{aligned} 1-\alpha &= P(-t_{\alpha/2,n-1} < T < t_{\alpha/2,n-1}) \\ &= P\left(-t_{\alpha/2,n-1} < \frac{\overline{X}-\mu}{\sqrt{S^2/n}} < t_{\alpha/2,n-1}\right) \end{aligned} \tag{2.11}$$

Doing a bit of algebra, we can see that

$$\frac{\overline{X}-\mu}{\sqrt{S^2/n}} < t_{\alpha/2,n-1} \tag{2.12}$$

is equivalent to

$$\overline{X} - t_{\alpha/2,n-1}\sqrt{\frac{S^2}{n}} < \mu \tag{2.13}$$

and

$$-t_{\alpha/2,n-1} < \frac{\overline{X}-\mu}{\sqrt{S^2/n}} \tag{2.14}$$

is equivalent to

$$\mu < \overline{X} + t_{\alpha/2,n-1}\sqrt{\frac{S^2}{n}} \tag{2.15}$$

so

$$P\left(\overline{X} - t_{\alpha/2,n-1}\sqrt{\frac{S^2}{n}} < \mu < \overline{X} + t_{\alpha/2,n-1}\sqrt{\frac{S^2}{n}}\right) = 1-\alpha \tag{2.16}$$

which defines a confidence interval for μ.

Tuesday 17 November 2015 – Read Section 4.4 of Hogg

Note that we'll skip Section 4.3, "Confidence Intervals for Parameters of Discrete Distributions".

3 Order Statistics

Order statistics are a way of formalizing and making more precise some of the descriptive statistics/exploratory data analysis techniques you may have used in the past. Recall, for example, that if you took a data sample and happened to get the values

x_1	x_2	x_3	x_4	x_5
17	6	19	3	12

you could sort them in order $\{3, 6, 12, 17, 19\}$ and then the median of the sample was the middle entry, 12. As a formal prescription, we could define y_1 to be the lowest value in the sample y_2 to be the next lowest, etc., so that you get

y_1	y_2	y_3	y_4	y_5
3	6	12	17	19

We can assign the same prescription to a random sample $\mathbf{X} = \{X_i\}$ drawn from some distribution with pdf $f(x)$. We define Y_1 to be the lowest value in the sample, Y_2 to be the next lowest, and so forth, up to Y_n, which is the highest. This means that (for a sample drawn from a continuous distribution), the statistics $\{Y_i\}$ will obey

$$Y_1 < Y_2 < \cdots < Y_n \tag{3.1}$$

with 100% probability. We note a couple of important facts:

- Each order statistic Y_j depends on the entire sample, i.e., all of the random variables $\{X_i\}$. For instance, in the realization above, changing x_4 to 20 would change y_1 to 6, y_2 to 12, etc.
- Although the random variables $\{X_i\}$ are independent, the $\{Y_i\}$ are not.

3.1 Joint pdf

We can write the joint pdf for the $\{Y_i\}$ in terms of the distribution from which the sample is drawn. To see how this works, consider the simple case where $n = 2$, so that

$$Y_1 = \min(X_1, X_2) \qquad \text{and} \qquad Y_2 = \max(X_1, X_2) \tag{3.2}$$

We know the joint pdf for X_1 and X_2 is

$$f_{\mathbf{X}}(x_1, x_2) = f(x_1)f(x_2) \tag{3.3}$$

The transformation between $\mathbf{X}$ and $\mathbf{Y}$ is a little different than what we've considered before, since it is not invertable. I.e., given values of y_1 and y_2, there is not a unique pair of x_1 and x_2 that you can determine corresponding to them. The pair $(y_1, y_2) = (3, 5)$ could correspond to either $(x_1, x_2) = (3, 5)$ or $(x_1, x_2) = (5, 3)$. (Note that the pair $(y_1, y_2) = (5, 3)$ is impossible by definition, so the joint pdf $f_{\mathbf{Y}}(y_1, y_2)$ must be zero unless $y_1 \le y_2$.) So to get the probability density at a particular pair of values (y_1, y_2) which satisfy $y_1 < y_2$ you need to add the probabilities for the two cases:

1. If $x_1 > x_2$, then $y_1 = x_2$ and $y_2 = x_1$; the Jacobian matrix associated with the transformation in this part of the (x_1, x_2) plane is

$$\left\{\frac{\partial x_i}{\partial y_j}\right\} = \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{3.4}$$

   Its determinant is $J = -1$ so $|J| = 1$ and the contribution to $f_{\mathbf{Y}}(y_1, y_2)$ from this case is $f_{\mathbf{X}}(y_2, y_1) = f(y_2)f(y_1)$.
2. If $x_1 < x_2$, then $y_1 = x_1$ and $y_2 = x_2$; the Jacobian matrix associated with the transformation in this part of the (x_1, x_2) plane is

$$\left\{\frac{\partial x_i}{\partial y_j}\right\} = \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{3.5}$$

   Its determinant is $J = 1$ so $|J| = 1$ and the contribution to $f_{\mathbf{Y}}(y_1, y_2)$ from this case is $f_{\mathbf{X}}(y_1, y_2) = f(y_1)f(y_2)$.

This means the total pdf is

$$f_{\mathbf{Y}}(y_1, y_2) = f(y_2)f(y_1) + f(y_1)f(y_2) = 2f(y_1)f(y_2) \qquad y_1 < y_2 \tag{3.6}$$

In the more general case of a sample of size n, there are $n!$ different ways the $\{X_i\}$ could be ordered, so there will be $n!$ equal contributions to $f_{\mathbf{Y}}(y_1, y_2, \ldots, y_n)$, and

$$f_{\mathbf{Y}}(y_1, y_2, \ldots, y_n) = n!\, f(y_1)f(y_2)\cdots f(y_n), \qquad y_1 < y_2 < \cdots < y_n \tag{3.7}$$

Note that

1. The restriction $y_1 < y_2 < \cdots < y_n$ is in addition to any other limitations on the support space associated with the distribution $f(x)$. (E.g., the support space for the order statistics from a sample drawn from an exponential distribution is $0 < y_1 < y_2 < \cdots < y_n < \infty$.)
2. Although the product $n!\, f(y_1)f(y_2)\cdots f(y_n)$ factors, the random variables $\{Y_i\}$ are not independent, because $y_1 < y_2 < \cdots < y_n$ is not a product space.

3.2 Quantiles

We've noted that, if n is odd, the order statistic $Y_{\frac{n+1}{2}}$ is the sample median. It can be used as an estimator of the median $\widetilde{\mu}$ of the probability distribution $f(x)$, which is defined by $F(\widetilde{\mu}) = \frac{1}{2}$, where $F(x)$ is the cumulative distribution function corresponding to $f(x)$, i.e.,

$$\int_{-\infty}^{\widetilde{\mu}} f(x)\,dx = 0.5 = \int_{\widetilde{\mu}}^{\infty} f(x)\,dx \tag{3.8}$$

This is a special case of a *quantile* ξ_p, defined by

$$F(\xi_p) = \int_{-\infty}^{\xi_p} f(x)\,dx = p \tag{3.9}$$

where $0 < p < 1$. To follow the notation of Hogg, we'll refer to the distribution median as $\xi_{0.5}$ rather than $\widetilde{\mu}$.

The order statistics can be used as estimators of quantiles of the distribution, and are referred to as *sample quantiles*. The convention adopted by Hogg is to divide the interval from 0 to 1 into $n+1$ equal pieces, and thus define Y_k as the $\frac{k}{n+1}$ sample quantile. Note that this convention is not unique; for instance, Devore uses $\frac{k-0.5}{n}$ in his section on probability plots. This is equivalent to dividing the interval $[0,1]$ into $n-1$ full-sized pieces and 2 half-sized pieces on the end. To see the difference between these prescriptions, consider the case $n = 4$:

k	1	2	3	4
$\frac{k}{n+1}$	0.2	0.4	0.6	0.8
$\frac{k-0.5}{n}$	0.125	0.375	0.625	0.875

3.2.1 q-q Plots

One way of checking whether a particular data set could reasonably be a sample drawn from a distribution is to compare the sample quantiles $\{y_k\}$ with the corresponding quantiles $\{\xi_{\frac{k}{n+1}}\}$ of the proposed distribution. This can be done on a plot with n points, $(\xi_{\frac{k}{n+1}}, y_k)$ for $k = 1, 2, \ldots, n$. If the points on this plot (known as a *probability plot* or q-q (for quantile-quantile) plot) are close to the line $y = \xi$, the sample is consistent with being drawn from the proposed distribution. A few things to note about this construction:

1. The horizontal components $\{\xi_{\frac{k}{n+1}}\}$ are determined by the size of the sample and the proposed distribution. They don't depend on the sample values.

2. This is only a qualitative construction. We don't have a rule for deciding whether the plot is "close enough" to the line or not.

3. You might think we'd need to make a plot for every possible distribution, but there are some families for which one plot will do for every member of the family. For instance, if the distribution in question is a normal distribution $N(\mu, \sigma^2)$ then the quantiles are $\{\xi_p\}$ defined by

$$p = F_X(\xi_p) = P(X \le \xi_p) = \Phi\left(\frac{\xi_p - \mu}{\sigma}\right) \tag{3.10}$$

where X is a $N(\mu, \sigma^2)$ random variable and $\Phi(z) = P(Z \le z)$ is the cdf for a standard normal random variable Z. But we also know that the quantiles of Z are given by

$$\Phi(\xi_{Z,p}) = p \tag{3.11}$$

(Note that in terms of the definitions used in constructing confidence intervals, $\xi_{Z,p} = z_{1-p}$ and that z_α is tabulated for various values of α.) Since the function $\Phi(z)$ is invertable,

$$\frac{\xi_p - \mu}{\sigma} = \xi_{Z,p} \tag{3.12}$$

so

$$\xi_p = \mu + \sigma\,\xi_{Z,p} \tag{3.13}$$

This means that if we plot y_k versus $\xi_{Z,\frac{k}{n+1}} = z_{\frac{n+1-k}{n+1}}$, then if $\{y_k\}$ are the order statistics of the realization of a sample drawn from a $N(\mu.\sigma^2)$ distribution, the points should lie close to the line $y = \mu + \sigma\xi$ where σ is the slope and μ is the intercept. This trick works for any family of distributions with only a scale and/or location parameter, where the cdf $F(x)$ is a function of $\frac{x-a}{b}$.

3.2.2 Confidence Intervals for Quantiles

Finally, consider a construction which allows us to make a confidence interval for one of the quantiles of an unknown distribution. This really is a nonparametric method, since while you could consider the unknown quantile to be the parameter, the underlying distribution cannot be determined by this parameter alone.

For simplicity, consider the special case of the median $\xi_{0.5}$, and a sample of size $n = 5$. We already know that the order statistic Y_3, which is the sample median, is a good estimator for $\xi_{0.5}$. Suppose we construct a confidence interval with Y_2 and Y_4 as endpoints. What is the confidence level associated with this? I.e., what is

$$P(Y_2 < \xi_{0.5} < Y_4) \; ? \tag{3.14}$$

Well, we know that by definition, the probability that any particular X_i is below the median $\xi_{0.5}$ is $P(X_i{<}\xi_{0.5}) = 0.5$. Since the random variables in the sample are independent, the question of whether each one is below the median is like an independent Bernoulli trial, which means the number of values in the sample below $\xi_{0.5}$ is a binomial random variable with $n = 5$ and $p = 0.5$. Now, if $Y_2 < \xi_{0.5}$ that means that at least two random variables in the sample are below the median. If $\xi_{0.5} < Y_4$ that less than four random variables in the sample are below the median. So $P(Y_2 < \xi_{0.5} < Y_4)$ is the probability that a $\text{Bin}(5, 0.5)$ random variable is equal to 2 or 3, i.e.,

$$\begin{aligned} P(Y_2 < \xi_{0.5} < Y_4) &= \binom{5}{2}(0.5)^2(0.5)^3 + \binom{5}{3}(0.5)^3(0.5)^2 \\ &= 2\left(\frac{10}{2^5}\right) = \frac{5}{8} = 0.625 \end{aligned} \tag{3.15}$$

So this is a confidence interval with confidence 0.625. In general, for a sample of size n, we can make a confidence interval on ξ_p using

$$P(Y_i < \xi_p < Y_j) = \sum_{k=i}^{j-1} \binom{n}{k} p^k (1-p)^{n-k} \tag{3.16}$$

This is a more limited construction than the usual confidence interval, since it's only possible for specific confidence levels determined by the binomial distribution, but for large n they are in practice pretty close together.

Thursday 19 November 2015 – Read Sections 4.5-4.6 of Hogg

4 Hypothesis Testing

So far, we've considered methods to get a handle on the unknown parameter(s) θ of a probability distribution $f(x;\theta)$ given that we draw a sample $\mathbf{X}$ from that distribution, with joint pdf

$$f_{\mathbf{X}}(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta) \tag{4.1}$$

and find a particular realization $\mathbf{X} = \mathbf{x}$. Now we want to consider how to use the realization of the sample to distinguish between two competing hypotheses about what the underlying distribution $f(x)$ is. In principle the differences could be qualitative, but for simplicity we'll assume that there is one family $f(x;\theta)$ parametrized by θ which lies somewhere in a region Ω and then take the hypotheses to be:

- H_0: the distribution is $f(x;\theta)$ where $\theta \in \omega_0$.
- H_1: the distribution is $f(x;\theta)$ where $\theta \in \omega_1$.

Typically, H_0 represents the absence of the effect we're looking for, and is known as the *null hypothesis*, while H_1 represents the presence of the effect, and is known as the *alternative hypothesis*.

For example, suppose someone claims to have extrasensory perception, and to be able to use their telepathic powers to determine the suits of cards drawn from a deck. For simplicity, assume we shuffle the deck after each draw. Then the data $\{X_i\}$ are a sample drawn from a Bernoulli distribution, with each X_i having some probability θ of being correct. The null hypothesis H_0 is that the person does not have ESP, and has a 25% chance of guessing each suit correctly, so $\theta = 0.25$. The alternative hypothesis H_1 is that they can determine the suit more accurately than by random chance (but perhaps not perfectly), so $\theta > 0.25$.

As another example, suppose that someone claims that when twins are born, the birth weight of the first twin is on average greater than that of the second. We could take the data $\{X_i\}$ to be the difference between the birth weights of the two twins, and assume that the weights are normally distributed with unknown variance. Then the null hypothesis H_0 is that $f(x)$ is a normal distribution with mean $\mu = 0$ and standard deviation $\sigma > 0$, while the alternative hypothesis H_1 is that $f(x)$ is a normal distribution with mean $\mu > 0$ and standard deviation $\sigma > 0$. (In this case there is a vector of parameters $\boldsymbol{\theta} = (\mu, \sigma)$.

A hypothesis test is simply a rule for choosing between the two hypotheses depending on the realization $\mathbf{x}$ of the sample $\mathbf{X}$. Stated most generally, we construct a critical region C which is a subset of the n-dimensional sample space $\mathcal{D}$. If $\mathbf{X} \in C$, we "reject the null hypothesis H_0", i.e., we favor H_1. If $\mathbf{X} \notin C$, i.e., $\mathbf{X} \in C^c$ we "accept the null hypothesis H_0", i.e., we favor H_0 over H_1. Now of course, since $\mathbf{X}$ is random, there will be some probability $P(\mathbf{X} \in C; \theta)$ that we'll reject the null hypothesis, which depends on the value of θ. If the test were perfect, that probability would be 0 if H_0 were true, i.e., for any $\theta \in \omega_0$, and 1 if H_1 were true, i.e., for any $\theta \in \omega_1$, but then we wouldn't be doing statistics. So instead there is some chance we will choose the "wrong" hypothesis, i.e., some probability that, given a value of $\theta \in \omega_0$ associated with H_0, the realization of our data will cause us to reject H_0, and some probability that, given a value of $\theta \in \omega_1$ associated with H_1, the realization of our data will cause us to accept H_0. As a bit of nomenclature,

- If H_0 is true and we reject H_0, this is called a *Type I Error* or a false positive.
- If H_1 is true and we reject H_0, we have made a correct decision (true positive).
- If H_0 is true and we accept H_0, we have made a correct decision (true negative).
- If H_1 is true and we accept H_0, this is called a *Type II Error* or a false negative.

Typically, a false positive is considered worse than a false negative, so usually we decide how high a false positive probability we can live with and then try to find the test which gives us the lowest false negative probability.

Given a critical region C, we'd like to talk about the associated false positive probability α and false negative probability $1 - \gamma$, but we have to be a bit careful, since H_0 and H_1 are in general *composite hypotheses*. This means that each of them corresponds not to a single parameter value θ and thus a single distribution, but rather to a range of values $\theta \in \omega_0$ or $\theta \in \omega_1$. So both α and γ may depend on the value of θ. We take the false alarm probability α to be the worst-case scenario within the null hypothesis

$$\alpha = \max_{\theta \in \omega_0} P(\mathbf{X} \in C; \theta) \tag{4.2}$$

This is also called the *size* of the critical region C. Somewhat confusingly, it's also referred to as the *significance* of the test. This is a bit counter intuitive, since a low value of α means the probability of a false positive is low, which means a positive

result is *more* significant than if α were higher. It is the probability that we'll falsely reject the null hypothesis H_0, maximized over any parameters within the range associated with H_0. On the other hand, since the alternative hypothesis almost always has a parameter θ associated with it, we define the probability of correctly rejecting the null hypothesis (which is one minus the probability of a false negative) as a function of θ:

$$\gamma_C(\theta) = P(\mathbf{X} \in C; \theta), \qquad \theta \in \omega_1 \tag{4.3}$$

We explicitly consider this as a function of the critical region C, since we might want to compare different tests with the same false alarm probability α (critical regions with the same size α) to see which is more powerful.

4.1 Example: Binomial Proportion

To give a concrete example, consider the ESP test described above. We let the would-be psychic predict the suit of n cards, count the total number of successes $Y = \sum_{i=1}^{n} X_i$, and reject the null hypothesis if $Y > k$ where k is some integer we've chosen, with $k > n/4$. For both of the hypotheses, Y is a binomial random variable, so

$$P(Y > k) = \sum_{i=k+1}^{n} \binom{n}{i} \theta^i (1-\theta)^{n-i} = 1 - F(k; \theta) \tag{4.4}$$

where

$$F(k; \theta) = \sum_{i=0}^{k} \binom{n}{i} \theta^i (1-\theta)^{n-i} \tag{4.5}$$

is the cdf of a binomial distribution $b(n, \theta)$. For the null hypothesis $\theta = 0.25$ and for the alternative hypothesis $0.25 < \theta < 1$. Thus the false alarm probability is

$$\alpha = 1 - F(k; 0.25) \tag{4.6}$$

and the power of the test is

$$\gamma_k(\theta) = 1 - F(k; \theta) \tag{4.7}$$

If we make the threshold k higher, we get a lower false alarm probability α, but we also get a less powerful test.

As a concrete example, suppose that $n = 20$, and we set a threshold of $k = 8$. We can use scipy, invoked by

```
ipython --pylab
```

to calculate the false alarm probability

```
In [1]: from scipy.stats import binom

In [2]: n = 20

In [3]: k = 8

In [4]: alpha = 1 - binom.cdf(k,n,0.25); alpha
Out[4]: 0.04092516770651855
```

So $\alpha \approx 0.041 = 4.1\%$. The power $\gamma(\theta)$ depends on the strength of the ESP effect, but suppose $\theta = 0.50$, that the psychic has a 1 in 2 chance rather than 1 in 4 of picking the right suit. Then we can calculate the power:

```
In [5]: gamma_50 = 1 - binom.cdf(k,n,0.50); gamma_50
Out[5]: 0.74827766418457031
```

so $\gamma(0.50) \approx 0.748 = 74.8\%$.

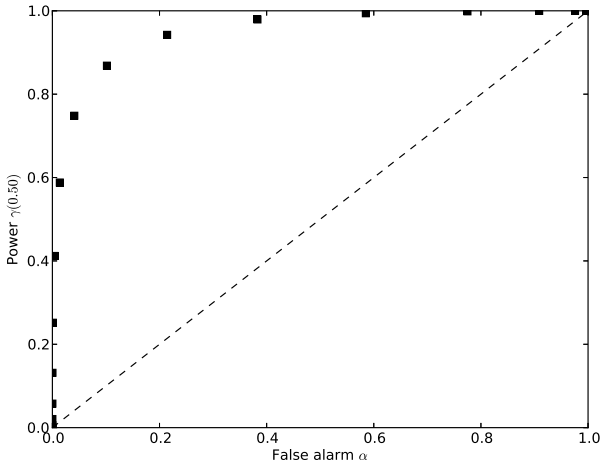
4.1.1 Aside: ROC Curves

We could make the test more powerful by lowering the threshold k, but then we would also increase the false alarm probability

α. A useful construction is the *receiver operating characteristic* curve, or ROC curve for short. Given a value of θ, we plot α versus $\gamma(\theta)$ for a range of threshold values k. We can do this with matplotlib as well, using the `arange` function to generate an array of integer values for k between 0 and 19:

```
In [6]: k = arange(20)

In [7]: alpha = 1 - binom.cdf(k,n,0.25)

In [8]: gamma_50 = 1 - binom.cdf(k,n,0.50)

In [9]: plot(alpha,gamma_50,'ks');

In [10]: xlabel(r'False alarm $\alpha$');

In [11]: ylabel(r'Power $\gamma(0.50)$');

In [12]: plot([0,1],[0,1],'k--');

In [13]: savefig('notes04_roc.eps');
```

The plot looks like this:



The diagonal line is $\gamma = \alpha$; we don't expect any sensible test to lie below this line, since it would mean that we were more likely to reject H_0 when it's true than when H_1 is true!

4.2 Example: Mean of a Normal Distribution

Consider the second example, where $\mathbf{X}$ is a random sample of size n from a normal distribution, where the null hypothesis H_0 is $\mu = 0$ and the alternative hypothesis H_1 is $\mu > 0$. For simplicity, let's assume that the variance σ^2 is actually known. (If the sample is large enough, we can use the sample variance s^2 as an estimate.) From our work on confidence intervals, we know that

$$P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} > z_\alpha\right) = \alpha \tag{4.8}$$

So if we define a critical region

$$C \equiv \frac{\overline{X}}{\sigma/\sqrt{n}} > z_\alpha \tag{4.9}$$

this will correspond to a test with false alarm rate α. The power of the test for a given true value of μ is

$$\begin{aligned}\gamma(\mu) &= P\left(\frac{\overline{X}}{\sigma/\sqrt{n}} > z_\alpha\right) = P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} > z_\alpha - \frac{\mu}{\sigma/\sqrt{n}}\right) \\ &= 1-\Phi\left(z_\alpha - \frac{\mu}{\sigma/\sqrt{n}}\right) = \Phi\left(\frac{\mu}{\sigma/\sqrt{n}} - z_\alpha\right)\end{aligned} \tag{4.10}$$

4.2.1 p-Values

In this example, as in the last one, we actually have a family of tests, parametrized by a threshold which we could imagine varying. Given a data realization $\mathbf{x}$, and in particular a sample mean $\overline{x}$, we will reject H_0 if $\overline{x} > z_\alpha\sigma/\sqrt{n}$. This means there will be some values of the false alarm probability α for which we reject H_0, and some for which we do not. One convenient way to report which tests would indicate a positive result (reject the null hypothesis) is to quote the α of the most stringent test for which H_0 would be rejected. Put another way, we ask, given a measurement (in this case $\overline{x}$), how likely is it that we would find a measurement at least this extreme, just by accident, if the null hypothesis were true. This is known as the p-value, and in this case it is defined as

$$p = P(\overline{X} \geq \overline{x}; \mu = 0) = 1-\Phi\left(\frac{\overline{x}}{\sigma/\sqrt{n}}\right) = \Phi\left(-\frac{\overline{x}}{\sigma/\sqrt{n}}\right) \tag{4.11}$$

A lower p value means that the results were less likely to have occurred by chance in the absence of a real effect (i.e., if the null hypothesis H_0 were true). Typically if $p < 0.05$, the result is considered interesting and worth future study.[3]

[3]However, if we test for many different effects, or test many different data sets, and only report the result with the lowest p value, we can greatly overstate the significance of our results. See `http://xkcd.com/882/`.

Note that the p value is often misinterpreted. It does *not* represent the probability that the null hypothesis is true (we cannot evaluate such a probability in frequentist inference). A p value of 0.01 simply means, for the statistic we decided to measure, if we repeated the test on many systems for which the null hypothesis was true, we'd get a measurement as extreme, or more, as the one we got, one percent of the time.

4.3 Aside: Bayesian Hypothesis Testing

In the Bayesian framework, the joint pdf for a sample $\mathbf{X}$ drawn from a distribution with parameter θ can be written as $f(\mathbf{x} \mid \theta)$. And we can talk about the posterior probability $P(H_1 \mid \mathbf{x})$ that a hypothesis is true, given an observed sample $\mathbf{x}$. There are a few challenges, though:

- Typically, our hypothesis allows us to describe the joint pdf for a random sample $\mathbf{X}$ collected in the presence of that hypothesis, $f(\mathbf{x} \mid H_1)$, and then we can use Bayes's theorem to construct the desired probability

$$P(H_1 \mid \mathbf{x}) = \frac{f(\mathbf{x} \mid H_1)\, P(H_1)}{f(\mathbf{x})} \tag{4.12}$$

  There are a couple of problematic quantities in this expression. First, it depends on $P(H_1)$, which is the prior probability that the hypothesis H_1 was true, which we might have difficulty stating. Second, the denominator $f(\mathbf{x})$ is the overall probability density for the sample, marginalized over a complete set of mutually exclusive models, i.e.,

$$\begin{aligned}f(\mathbf{x}) = f(\mathbf{x} \mid H_0)\, P(H_0) + f(\mathbf{x} \mid H_1)\, P(H_1) \\ + f(\mathbf{x} \mid H_2)\, P(H_2) + \cdots\end{aligned} \tag{4.13}$$

  which we're even less likely to have a handle on. The usual way around this is to calculate not $P(H_1 \mid \mathbf{x})$, but rather

the *odds ratio*, the ratio of the posterior probabilities of the competing models H_1 and H_0.

$$\frac{P(\,H_1 \mid \mathbf{x}\,)}{P(\,H_0 \mid \mathbf{x}\,)} = \frac{f(\,\mathbf{x} \mid H_1\,)\,P(H_1)/\cancel{f(\mathbf{x})}}{f(\,\mathbf{x} \mid H_0\,)\,P(H_0)/\cancel{f(\mathbf{x})}} = \frac{f(\,\mathbf{x} \mid H_1\,)}{f(\,\mathbf{x} \mid H_0\,)}\frac{P(H_1)}{P(H_0)} \tag{4.14}$$

We still have the ratio of the prior probabilities, $P(H_1)/P(H_0)$, but can at least calculate unambiguously the factor

$$\frac{f(\,\mathbf{x} \mid H_1\,)}{f(\,\mathbf{x} \mid H_0\,)} \tag{4.15}$$

by which we modify the prior odds ratio to get the posterior one. This quantity is known as the *Bayes factor*.

- The joint pdfs $f(\,\mathbf{x} \mid H_1\,)$ and $f(\,\mathbf{x} \mid H_0\,)$ require that we specify the hypotheses H_1 and H_0 a little more precisely than we've done so far. In particular, if one or both of them is a composite hypothesis, it's not good enough to specify a range like $\theta \in \omega_1$ for the parameter(s) θ. We need to say what the probability density associated with the hypothesis for θ. For instance, marginalizing over θ gives

$$f(\,\mathbf{x} \mid H_1\,) = \int_{\omega_1} f(\,\mathbf{x} \mid \theta\,)\,f(\,\theta \mid H_1\,)\,d\theta \tag{4.16}$$

Still, if we can overcome these mostly technical hurdles, the Bayesian methods are very useful and satisfying. For example, once we specify the hypotheses, including any prior distributions of the parameters, the construction of the Bayes factor $f(\,\mathbf{x} \mid H_1\,)/f(\,\mathbf{x} \mid H_0\,)$ from the data sample $\mathbf{x}$ is unique. In contrast, in classical (frequentist) hypothesis testing, we have to choose a test or family of tests, which usually means boiling the random sample $\mathbf{X}$ down to a single statistic and hoping we've kept the right information and haven't discarded something that could have made the test more sensitive.

Tuesday 24 November 2015 – Read Section 4.7 of Hogg

4.4 Chi-Square Tests

Recall that if we have a statistical test that tells us to favor an alternative hypothesis H_1 over the null hypothesis H_0 if a random sample $\mathbf{X}$ falls in a critical region C, the false alarm probability for the test (also known as the size of C or the significance of the test) is

$$\alpha = P(\mathbf{X} \in C|H_0) \tag{4.17}$$

Similarly, if we have some statistic $Y = q(\mathbf{X})$ which tends to be higher if the data favor H_0, rather than fixing a threshold y_0 and defining $\mathbf{X} \in C$ to mean $q(\mathbf{X}) > y_0$, we can use the actual observed value of the statistic $y = q(\mathbf{x})$ to define the p-value

$$p = P(Y > y|H_0) \tag{4.18}$$

i.e., the probability of getting a result at least this extreme if H_0 is the correct hypothesis.

Notice that neither of these constructions (for α or p) depend on the alternative hypothesis H_1 at all; they're simply ways of talking about how *inconsistent* the data are with the null hypothesis H_0. In the same category is the so-called goodness-of-fit test: we measure how closely the actually-collected data come to the most likely values, and how likely we were to deviate by that much or more, if the model were actually correct.

We want to consider a family of goodness-of-fit tests related to the chi-square distribution. We'll generalize the picture a bit to allow the random vector $\mathbf{X}$ predicted by the model to be not just a random sample, but any set of n independent (but not identically distributed) random variables. In several cases we can construct a statistic which is, exactly or approximately, chi-squared distributed. As an overview, we will consider four cases:

- If each X_i obeys a $N(\mu_i, \sigma_i^2)$ distribution, we know $\sum_{i=1}^{n} \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2$ obeys a $\chi^2(n)$ distribution.
- Given a binomial random variable, we can construct a statistic which is approximately $\chi^2(1)$-distributed, if the number of trials is large.
- We will generalize this to a multinomial random variable with k alternatives to get a $\chi^2(k-1)$-distributed statistic. (We won't show this.)
- If the model we're testing has m parameters, and we pick those to give the best fit (minimum chi-square statistic) given our n data points, the minimized statistic will be approximately $\chi^2(n-m)$-distributed. (We won't show this, but it makes sense, especially if the modelled means are linear in the parameters.)

First, consider the case where the null hypothesis tells us that our random data vector $\mathbf{X}$ is made up of n independent random variables and that X_i is $N(\mu_i, \sigma_i^2)$ with some specified μ_i and σ_i. Then we know that

$$Z_i = \frac{X_i - \mu_i}{\sigma_i} \tag{4.19}$$

is a standard normal random variable, and that

$$Y = \sum_{i=1}^{n} (Z_i)^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2 \tag{4.20}$$

being the sum the squares of of n independent standard normal random variables, is a $\chi^2(n)$ random variable. We can use the cdf of the χ^2 distribution to find the p value associated with a particular mismatch. In particular

$$P(Y > \chi^2_{n,0.05}) = 0.05 \tag{4.21}$$

where $\chi^2_{n,0.05}$ is the 95th percentile of the $\chi^2(n)$ distribution.

4.4.1 Binomial and multinomial experiments

Even if the random variables are not normally distributed, a statistic constructed in this way may still be approximately χ^2-distributed. For example, suppose that X_1 is a $b(n, p_1)$ binomial random variable. We know that $E(X_1) = np_1$ and $\mathrm{Var}(X_1) = np_1(1-p_1)$. We also know that if np_1 and $n(1-p_1)$ are both more than around 5, the probabilities associated with the binomial distribution can be approximated using the corresponding normal distribution, $N(np_1, np_1[1-p_1])$. In that case, we get an approximate $\chi^2(1)$ random variable

$$Q = \frac{(X_1 - np_1)^2}{np_1(1-p_1)} \tag{4.22}$$

We've been putting on the subscript $_1$ to allow a change of perspective. If we think of the binomial distribution as a special case of the multinomial with $k = 2$, then we also have probability $p_2 = 1 - p_1$ and a dependent random variable $X_2 = n - X_1$. The statistic can then be written as

$$Q = \frac{(X_1 - np_1)^2}{np_1p_2} \tag{4.23}$$

If we notice that $\frac{1}{p_1} + \frac{1}{p_2} = \frac{p_1+p_1}{p_1p_2} = \frac{1}{p_1p_2}$ and $X_2 - np_2 = n - X_1 - n(1-p_1) = X_1 - np_1$, we see we can write the statistic more symmetrically as

$$Q = \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2} \tag{4.24}$$

It turns out that, subject to the same requirements of approximate normality ($np_i > 5$), for a multinomial distribution with k alternatives, the analogous statistic

$$Q = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} \tag{4.25}$$

is approximately $\chi^2(k-1)$ distributed. This can be used to, for example, test a die to see if it's fair, e.g., by rolling a six-sided die more than about 30 times and counting up the number of ones, twos, etc.

4.4.2 Minimizing over parameters

Finally, consider the case where H_0 is a composite hypothesis, corresponding to a family of models, and says that X_i is $N(\mu_i(\boldsymbol{\theta}), \sigma_i^2(\boldsymbol{\theta}))$ where $\boldsymbol{\theta}$ is an m-dimensional vector of parameters $\theta_1, \ldots, \theta_m$. Then the χ^2 statistic corresponding to a particular set of parameter values $\boldsymbol{\theta}$ is

$$Y(\boldsymbol{\theta}) = q(\mathbf{X}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \left(\frac{X_i - \mu_i(\boldsymbol{\theta})}{\sigma_i(\boldsymbol{\theta})} \right)^2 \tag{4.26}$$

If H_0 is true, and $\boldsymbol{\theta}$ are the actual parameter values, the statistic $Y(\boldsymbol{\theta})$ is $\chi^2(n)$-distributed. Suppose that we don't know the parameter values, though, and we've collected a data vector $\mathbf{x}$. We can choose as our best estimate of the parameters the values $\widehat{\boldsymbol{\theta}}$ which minimize the chi-square statistic $y(\boldsymbol{\theta}) = q(\mathbf{x}, \boldsymbol{\theta})$, i.e., $\widehat{\boldsymbol{\theta}}(\mathbf{x})$ is the solution to the system of equations

$$\widehat{\boldsymbol{\theta}}(\mathbf{x}) \text{ satisfies} \qquad \frac{\partial y(\boldsymbol{\theta})}{\partial \theta_j} = \frac{\partial q(\mathbf{x}, \boldsymbol{\theta})}{\partial \theta_j} = 0, \quad j = 1, \ldots, m \tag{4.27}$$

If we put these values back into $y(\boldsymbol{\theta})$, we get the minimized chi-square value $\widehat{y} = y(\widehat{\boldsymbol{\theta}}) = q(\mathbf{x}, \widehat{\boldsymbol{\theta}}(\mathbf{x}))$. Using this prescription to generate a statistic, we have

$$\widehat{Y} = Y(\widehat{\boldsymbol{\theta}}(\mathbf{X})) = q(\mathbf{X}, \widehat{\boldsymbol{\theta}}(\mathbf{X})) \tag{4.28}$$

It turns out that, under many circumstances, the minimized chi-square statistic $\widehat{Y}$ obeys a $\chi^2(n-m)$ distribution. We won't prove this, but note that it can be shown exactly for the case where the variances $\{\sigma_i^2\}$ are independent of the parameters $\boldsymbol{\theta}$ and the means $\{\mu_i\}$ depend linearly on the parameters, $\boldsymbol{\mu}(\boldsymbol{\theta}) = \mathbf{B}\boldsymbol{\theta}$ where $\mathbf{B}$ is a $n \times m$ matrix of rank $m < n$. See for example section 5.6 of `http://ccrg.rit.edu/~whelan/courses/2014_1sp_ASTP_611/notes_probability.pdf`.

To illustrate this with a concrete example, suppose that H_0 says that, for a parameter θ, the data are a random sample drawn from a $N(\theta, \sigma)$ distribution, i.e., each of the μ_i is equal to θ and each of the σ_i is equal to σ. Then the chi-square statistic is

$$Y(\theta) = \sum_{i=1}^{n} \frac{(X_i - \theta)^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \theta)^2 \tag{4.29}$$

For a particular data realization, $y(\theta) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \theta)^2$ and the minimum is found by solving

$$0 = y'(\widehat{\theta}) = 2\frac{1}{\sigma^2} \sum_{i=1}^{n} (\widehat{\theta} - x_i) = 2\frac{1}{\sigma^2} \left(n\widehat{\theta} - \sum_{i=1}^{n} x_i \right) \tag{4.30}$$

I.e., $\widehat{\theta}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} x_i$ and the corresponding estimator is

$$\widehat{\theta}(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^{n} X_i = \overline{X} \tag{4.31}$$

i.e., the sample mean. Note that $\widehat{\theta}$ is not only the value which maximizes the chi-square statistic, it's also the maximum likelihood estimate. This is because the likelihood function is

$$\begin{aligned} L(\theta) = f_{\mathbf{X}}(\mathbf{x}, \theta) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \frac{(x_i - \theta)^2}{\sigma^2} \right) \\ &= (2\pi\sigma^2)^{-n/2} e^{-y/2} \propto e^{-q(\mathbf{x},\theta)/2} \end{aligned} \tag{4.32}$$

I.e., the likelihood is a constant times the exponential of minus one-half the χ^2 value. This will be true as long as the underlying

distribution is normal and the variances don't depend on the parameters (so that the constant out front really is a constant).

Anyway, if we substitute the mle $\widehat{\theta}(\mathbf{X}) = \overline{X}$ back into the chi-square statistic, we get the minimized chi-square

$$\widehat{Y} = Y(\theta = \overline{X}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n-1}{\sigma^2}S^2 \qquad (4.33)$$

where S^2 is the sample variance. But this is a combination which we know to be $\chi^2(n-1)$-distributed as a result of Student's theorem, so we've confirmed in this case that the chi-square statistic minimized over the $m = 1$ parameters obeys a $\chi^2(n-m)$ distribution, as advertised.