PROBABILITY AND STATISTICS

MANJUNATH KRISHNAPUR

CONTENTS

Statistics

1. INTRODUCTION

In statistics we are faced with data, which could be measurements in an experiment, responses in a survey etc. There will be some randomness, which may be inherent in the problem or due to errors in measurement etc. The problem in statistics is to make various kinds of inferences about the underlying distribution, from realizations of the random variables. We shall consider a few basic types of problems encountered in statistics. We shall mostly deal with examples, but sufficiently many that the general ideas should become clear too. It may be remarked that we stay with the simplest "textbook type problems" but we shall also see some real data. Unfortunately we shall not touch upon the problems of current interest, which typically involve very huge data sets etc. Here are the kinds of problems we study.

General setting: We shall have data (measurements perhaps), usually of the form $X_1, \ldots, X_n$ which are realizations of independent random variables from a common distribution. The underlying distribution is not known. In the problems we consider, typically the distribution is known, except for the values of a few parameters. Thus, we may write the data as $X_1, \ldots, X_n$ i.i.d. $f_\theta(x)$ where $f_\theta(x)$ is a pdf or pmf for each value of the parameter(s) θ. For example, the density could be of $N(\mu, \sigma^2)$ (two unknown parameters μ and σ^2) or of Pois(λ) (one unknown parameter λ).

(1) Estimation: Here, the question is to guess the value of the unknown θ from the sample $X_1, \ldots, X_n$. For example, if X_i are i.i.d. from Ber(p) distribution (p is unknown), then a reasonable guess for θ would be the sample mean $\bar{X}_n$ (an *estimator*). Is this the only one? Is it the "best" one? Such questions are addressed in estimation.

(2) Confidence intervals: Here again the problem is of estimating the value of a parameter, but instead of giving one value as a guess, we instead give an interval and quantify how sure we are that the interval will contain the unknown parameter. For example, a coin with unknown probability p of turning up head, is tossed n times. Then, a confidence interval for p could be of the form

$$\left[\bar{X}_n - \frac{3}{\sqrt{n}}\sqrt{\bar{X}_n(1-\bar{X}_n)}, \bar{X}_n + \frac{3}{\sqrt{n}}\sqrt{\bar{X}_n(1-\bar{X}_n)}\right]$$

where $\bar{X}_n$ is the proportion of heads in n tosses. The reason for such an interval will come later. It turns out that if n is large, one can say that with probability 0.99 ("confidence level"), this interval will contain the true value of the parameter.

(3) Hypothesis testing: In this type of problem we are required to decide between two competing choices ("hypotheses"). For example, it is claimed that one batch of students is better than a second batch of students in mathematics. One way to check this is to give the same exam to students in both exams and record the scores. Based on the scores, we have to decide whether the first batch is

better than the second (one hypothesis) or whether there is not much difference between the two (the other hypothesis). One can imagine that this can be done by comparing the sample means etc., but that will come later.

A good analogy for testing problems is from law, where the judge has to decide whether an accused is guilty or not guilty. Evidence presented by lawyers take the role of data (but of course one does not really compute any probabilities quantitatively here!).

(4) Regression: Consider two measurements, such as height and weight. It is reasonable to say that weight and height are positively correlated (if the height is larger, the weight tends to be larger too), but is there a more quantitative relationship? Can we predict the weight (roughly) from the height? One could try to see if a linear function fits: wt. $= a$ ht. $+ b$ for some a, b. Or perhaps a more complicated fit such as wt. $= a$ ht. $+ b$ ht.$^2 + c$, etc. To see if this is a good fit, and to know what values of a, b, c to take, we need data. Thus, the problem is that we have some data (H_i, W_i), $i = 1, 2, \ldots, n$, and based on this data we try to find the best linear fit (or the best quadratic fit) etc.

As another example, consider the approximate law that the resistivity of a material is proportional to the temperature. What is the constant of proportionality (for a given material). Here we have a law that says $R = aT$ where a is not known. By taking many measurements at various temperatures we get data (T_i, R_i), $i = 1, 2, \ldots, n$. From this we must find the best possible a (if all the data points were to lie on a line $y = ax$, there would be no problem. In reality they never will, and that is why the choice is an issue!).

2. Estimation problems

Consider the following examples.

(1) A coin has an unknown probability p of turning up head. We wish to determine the value of p. For this, we toss the coin 100 times and observe the outcomes. How to give a guess for the value of p based on the data?

(2) A factory manufacture light bulbs whose lifetimes may be assumed to be exponential random variables with a mean life-time μ. We take a sample of 50 bulbs at random and measure their life-times $X_1, \ldots, X_{50}$. Based on this data, how can we present a reasonable guess for μ? We may want to do this so that the specifications can be printed on the product when sold.

(3) Can we guess the average height μ of all people in India by taking a random sample of 100 people and measuring their heights?

In such questions, there is an unknown parameter μ (there could be more than one unknown parameter too) whose value we are trying to guess based on the data. The data consists of i.i.d. random variables from a family of distributions. We assume that the family of distributions is

known and the only unknown is (are) the value of the parameter(s). Rather than present the ideas in abstract let us see a few examples.

Example 1. Let $X_1, \ldots, X_n$ be i.i.d. random variables with Exponential density $f_\mu(x) = \frac{1}{\mu} e^{-x/\mu}$ (fro $x > 0$) where the value of $\mu > 0$ is unknown. How to *estimate* it using the data $X = (X_1, \ldots, X_n)$?

This is the framework in which we would study the second example above, namely the lie-time distribution of light bulbs. Observe that we have parameterized the exponential family of distributions differently from usual. We could equivalently have considered $g_\lambda(x) = \lambda e^{-\lambda x}$ but the interest is then in estimating $1/\lambda$ (which is the expected value) rather than λ. Here are two methods.

Method of moments: We observe that $\mu = \mathbf{E}_\mu[X_1]$, the mean of the distribution (also called *population mean*). Hence it seems reasonable to take the sample mean $\bar{X}_n$ as an estimate. On second thought, we realize that $\mathbf{E}_\mu[X_1^2] = 2\mu^2$ and hence $\mu = \sqrt{\frac{1}{2}\mathbf{E}_\mu[X_1^2]}$. Therefore it also seems reasonable to take the corresponding sample quantity, $T_n := \sqrt{\frac{1}{2n}(X_1^2 + \ldots + X_n^2)}$ as an estimate for μ. One can go further and write μ in various ways as $\mu = \sqrt{\text{Var}_\mu(X_1)}$, $\mu = \sqrt[3]{\frac{1}{6}\mathbf{E}_\mu[X_1^3]}$ etc. Each such expression motivates an estimate, just by substituting sample moments for population moments.

This is called estimating by the *method of moments* because we are equating the sample moments to population moments to obtain the estimate.

We can also use other features of the distribution, such as quantiles (we may call this the "method of quantiles"). In other words, obtain estimates by equating the sample quantiles to population quantiles. For example, the median of X_1 is $\mu \log 2$, hence a reasonable estimate for μ is $M_n/\log 2$, where M_n is a sample median. Alternately, the 25% quantile of Exponential$(1/\mu)$ distribution is $\mu \log(4/3)$ and hence another estimate for μ is $Q_n/\log(4/3)$ where Q_n is a 25% sample quantile.

Maximum likelihood method: The joint density of $X_1, \ldots, X_n$ is

$$g_\mu(x_1, \ldots, x_n) = \mu^{-n} e^{-\mu(x_1 + \ldots + x_n)} \qquad \text{if all } x_i > 0$$

(since X_i are independent, the joint density is a product). We evaluate the joint density at the observed data values. This is called the likelihood function. In other words, define,

$$L_X(\mu) := \mu^{-n} e^{-\frac{1}{\mu}\sum_{i=1}^n X_i}.$$

Two points: This is the joint density of $X_1, \ldots, X_n$, evaluated at the observed data. Further, we like to think of it as a function of μ with $X := (X_1, \ldots, X_n)$ being fixed.

When μ is the actual value, then $L_X(\mu)$ is the "likelihood" of seeing the data that we have actually observed. The *maximum likelihood estimate* is that value of μ that maximizes the likelihood function. In our case, by differentiating and setting equal to zero we get,

$$0 = \frac{d}{d\mu} L_X(\mu) = -n\mu^{-n-1} e^{-\frac{1}{\mu}\sum_{i=1}^n X_i} + \mu^{-n} \left(\frac{1}{\mu^2} \sum_{i=1}^n X_i \right) e^{-\frac{1}{\mu}\sum_{i=1}^n X_i}$$

which is satisfied when $\mu = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}_n$. To distinguish this from the true value of μ which is unknown, it is customary to put a hat on the leter μ. We write $\hat{\mu}_{MLE} = \bar{X}_n$. We should really verify whether $L(\mu)$ is maximized or minimized (or neither) at this point, but we leave it to you to do the checking (eg., by looking at the second derivative).

Let us see the same methods at work in two more examples.

Example 2. Let $X_1, \ldots, X_n$ be i.i.d. Ber(p) random variables where the value of p is unknown. How to *estimate* it using the data $X = (X_1, \ldots, X_n)$?

Method of moments: We observe that $p = \mathbf{E}_p[X_1]$, the mean of the distribution (also called *population mean*). Hence, a method of moments estimator would be the sample mean $\bar{X}_n$. In this case, $\mathbf{E}_p[X_1^2] = p$ again but we don't get any new estimate because $X_k^2 = X_k$ (as X_k is 0 or 1)

Maximum likelihood method: Now we have a probability mass function instead of density. The joint pmf of of $X_1, \ldots, X_n$ is $f_p(x_1, \ldots, x_n = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}$ when each x_i is 0 or 1. The likelihood function is

$$L_X(p) := p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} = p^{n\bar{X}_n}(1-p)^{n(1-\bar{X}_n)}.$$

We need to find the value of p that maximizes $L_X(p)$. Here is a trick that almost always simplifies calculations (try it in the previous example too!). Instead of maximizing $L_X(p)$, maximize $\ell_X(p) = \log L_X(p)$ (called the *log-likelihood function*). Since "log" is an increasing function, the maximizer will remain the same. In our case,

$$\ell_X(p) = \bar{X}_n \log p + n(1-\bar{X}_n)\log(1-p).$$

Differentiating and setting equal to 0, we get $\hat{p}_{MLE} = \bar{X}_n$. Again the sample mean is the maximum likelihood estimate.

A last example.

Example 3. Consider the two-parameter Laplace-density $f_{\theta,\alpha}(x) = \frac{1}{2\alpha} e^{-\frac{|x-\theta|}{\alpha}}$ for all $x \in \mathbb{R}$. Check that $f_{\theta,\alpha}$ is indeed a density for all $\theta \in \mathbb{R}$ and $\alpha > 0$.

Now suppose we have data $X_1, \ldots, X_n$ i.i.d. from $f_{\theta,\alpha}$ where we do not know the values of θ and α. How to estimate the parameters?

Method of moments: We compute

$$\mathbf{E}_{\theta,\alpha}[X_1] = \frac{1}{2\alpha}\int_{-\infty}^{+\infty} te^{-\frac{|t-\theta|}{\alpha}}dt = \frac{1}{2}\int_{-\infty}^{+\infty}(\alpha s+\theta)e^{-|s|}ds = \theta.$$

$$\mathbf{E}_{\theta,\alpha}[X_1^2] = \frac{1}{2\alpha}\int_{-\infty}^{+\infty} t^2e^{-\frac{|t-\theta|}{\alpha}}dt = \frac{1}{2}\int_{-\infty}^{+\infty}(\alpha s+\theta)^2e^{-|s|}ds = 2\alpha^2+\theta^2.$$

Thus the variance is $\mathrm{Var}_{\theta,\alpha}(X_1) = 2\alpha^2$. Based on this, we can take the method of moments estimate to be $\hat{\theta}_n = \bar{X}_n$ (sample mean) and $\hat{\alpha}_n = \frac{1}{\sqrt{2}}s_n$ where $s_n^2 = \frac{1}{n-1}\sum_{i=1}^n(X_i-\bar{X}_n)^2$. At the moment the ideas of defining sample variance as s_n^2 may look strange and it might be more natural to take $V_n := \frac{1}{n}\sum_{i=1}^n(X_i-\bar{X}_n)^2$ as an estimate for the population variance. As we shall see later, s_n^2 has some desirable properties that V_n lacks. Whenever we say sample variance, we mean s_n^2, unless stated otherwise.

Maximum likelihood method: The likelihood function of the data is

$$L_X(\theta,\alpha) = \prod_{k=1}^n \frac{1}{2\alpha}\exp\left\{-\frac{|X_k-\theta|}{\alpha}\right\} = 2^{-n}\alpha^{-n}\exp\left\{-\sum_{k=1}^n\frac{|X_k-\theta|}{\alpha}\right\}.$$

The log-likelihood function is

$$\ell_X(\theta,\alpha) = \log L(\theta,\alpha) = -n\log 2 - n\log\alpha - \frac{1}{\alpha}\sum_{k=1}^n|X_k-\theta|.$$

We know that[1] for fixed $X_1,\ldots,X_n$, the value of $\sum_{k=1}^n|X_k-\theta|$ is minimized when $\theta = M_n$, the median of $X_1,\ldots,X_n$ (strictly speaking the median may have several choices, all of them are equally good). Thus we fix $\hat{\theta} = M_n$ and then we maximize $\ell(\hat{\theta},\alpha)$ over α by differentiating. We get $\hat{\alpha} = \frac{1}{n}\sum_{k=1}^n|X_k-\theta|$ (the sample mean-absolute deviation about the median). Thus the MLE of (θ,α) is $(\hat{\theta},\hat{\alpha})$.

In homeworks and tutorials you will see several other estimation problems which we list in the exercise below.

[1] If you do not know here is an argument. Let $x_1 < x_2 < \ldots < x_n$ be n distinct real numbers and let $a \in \mathbb{R}$. Rewrite $\sum_{k=1}^n|x_k-a|$ as $(|x_1-a|+|x_n-a|)+(|x_2-a|+|x_{n-1}-a|)+\ldots$. By triangle inequality, we see that

$$|x_1-a|+|x_n-a| \geq x_n-x_1, \quad |x_2-a|+|x_{n-1}-a| \geq x_{n-1}-x_2, \quad |x_3-a|+|x_{n-2}-a| \geq x_{n-2}-x_3\ldots.$$

Further the first inequality is an equality if and only if $x_1 \leq a \leq x_n$, the second inequality is an equality if and only if $x_2 \leq a \leq x_{n-1}$ etc. In particular, if a is a median, then all these inequalities become equalities and shows that a median minimizes the given sum.

Exercise 4. Find an estimate for the unknown parameters by the method of moments and the maximum likelihood method.

(1) $X_1, \ldots, X_n$ are i.i.d. $N(\mu, 1)$. Estimate μ. How do your estimates change if the distribution is $N(\mu, 2)$?

(2) $X_1, \ldots, X_n$ are i.i.d. $N(0, \sigma^2)$. Estimate σ^2. How do your estimates change if the distribution is $N(7, \sigma^2)$?

(3) $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$. Estimate μ and σ^2.

[**Note:** The first case is when σ^2 is known and μ is unknown. Then the known value of σ^2 may be used to estimate μ. In the second case it is similar, now μ is known and σ^2 is not known. In the third case, both are unknown].

Exercise 5. $X_1, \ldots, X_n$ are i.i.d. Geo(p) Estimate $\mu = 1/p$.

Exercise 6. $X_1, \ldots, X_n$ are i.i.d. Pois(λ) Estimate λ.

Exercise 7. $X_1, \ldots, X_n$ are i.i.d. Beta(a, b) Estimate a, b.

The following exercise is approachable by the same methods but requires you to think a little.

Exercise 8. $X_1, \ldots, X_n$ are i.i.d. Uniform$[a, b]$ Estimate a, b.

3. Properties of estimates

We have seen that there may be several competing estimates that can be used to estimate a parameter. How can one choose between these estimates? In this section we present some properties that may be considered desirable in an estimator. However, having these properties does not lead to an unambiguous choice of one estimate as the best for a problem.

The setting: Let $X_1, \ldots, X_n$ be i.i.d random variables with a common density $f_\theta(x)$. The parameter θ is unknown and the goal is to estimate it. Let T_n be an estimator for θ, this just means that T_n is a function of $X_1, \ldots, X_n$ (in words, if we have the data at hand, we should be able to compute the value of T_n).

Bias: Define the *bias* of the estimator as $\text{bias}_{T_n}(\theta) := \mathbf{E}_\theta[T_n] - \theta$. If $\text{Bias}_{T_n}(\theta) = 0$ for all values of the parameter θ then we say that T_n is *unbiased* for θ. Here we write θ in the subscript of $\mathbf{E}_\theta$ to

remind ourself that in computing the expectation we use the density f_θ. However we shall often omit the subscript for simplicity.

Mean-squared error: The *mean squared error* of T_n is defined as $\text{m.s.e.}_{T_n}(\theta) = \mathbf{E}_\theta[(T_n - \theta)^2]$. This is a function of θ. Smaller it is, better our estimate.

In computing mean squared error, it is useful to observe the formula

$$\text{m.s.e.}_{T_n}(\theta) = \text{Var}_{T_n}(\theta) + \left(\text{Bias}_{T_n}(\theta)\right)^2.$$

To prove this, consider and random variable Y with mean μ and observe that for any real number a we have

$$\begin{aligned}\mathbf{E}[(Y-a)^2] &= \mathbf{E}[(Y-\mu+\mu-a)^2] = \mathbf{E}[(Y-\mu)^2] + (\mu-a)^2 + 2(\mu-a)\mathbf{E}[Y-\mu] \\ &= \mathbf{E}[(Y-\mu)^2] + (\mu-a)^2 = \text{Var}(Y) + (\mu-a)^2.\end{aligned}$$

Use this identity with T_n in place of Y and θ in place of a.

Remark 9. An analogy. Consider shooting with a rifle having a telescopic sight. A given target can be missed for two reasons. One, the marksman may be unskilled and shoot all over the place, sometimes a meter to the right of the target, sometimes a meter to the left, etc. In this case, the shots have a large variance. Another person may consistently hit a point 20 cm. to the right of the target. Perhaps the telescopic sight is not set right, and this caused the systematic error. This is the bias. Both bias and variance contribute to missing the target.

Example 10. Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$. Let $V_n = \frac{1}{n}\sum_{k=1}^n (X_k - \bar{X}_n)^2$ be an estimate for σ^2. By expanding the squares we get

$$V_n = \bar{X}_n^2 + \frac{1}{n}\sum_{k=1}^n X_k^2 - \frac{2}{n}\bar{X}_n \sum_{k=1}^n X_k = \left(\frac{1}{n}\sum_{k=1}^n X_k^2\right) - \bar{X}_n^2.$$

It is given that $\mathbf{E}[X_k] = \mu$ and $\text{Var}(X_k) = \sigma^2$. Hence $\mathbf{E}[X_k^2] = \mu^2 + \sigma^2$. We have seen before that $\text{Var}(\bar{X}_n) = \sigma^2$ and $\mathbf{E}[\bar{X}_n] = \mu$. Hence $\mathbf{E}[\bar{X}_n^2] = \mu^2 + \frac{\sigma^2}{n}$. Putting all this together, we get

$$\mathbf{E}\left[V_n\right] = \left(\frac{1}{n}\sum_{k=1}^n \mu^2 + \sigma^2\right) - \left(\mu^2 + \frac{\sigma^2}{n}\right) = \frac{n-1}{n}\sigma^2.$$

Thus, the bias of V_n is $\frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{1}{n}\sigma^2$.

Example 11. For the same setting as the previous example, suppose $W_n = \frac{1}{n}\sum_{k=1}^n (X_k - \mu)^2$. Then it is easy to see that $\mathbf{E}[W_n] = \sigma^2$. Can we say that W_n is an unbiased estimate for σ^2? There is a hitch!

If the value of μ is unknown, then W_n is *not* an estimate (cannot compute it using $X_1, \ldots, X_n$!). However if μ is known, then it is an unbiased estimate. For example, if we knew that $\mu = 0$, then $W_n = \frac{1}{n}\sum_{k=1}^n X_k^2$ is an unbiased estimate for σ^2.

When μ is unknown, we define $s_n^2 = \frac{1}{n-1}\sum_{k=1}^n (X_k - \bar{X}_n)^2$. Clearly $s_n^2 = \frac{n}{n-1}V_n$ and hence $\mathbf{E}[s_n^2] = \frac{n}{n-1}\mathbf{E}[V_n] = \sigma^2$. Thus, s_n^2 is an unbiased estimate for σ^2. Note that s_n^2 depends only on the data and hence it is an estimate, whether μ is known or unknown.

All the remarks in the above two examples apply for any distribution, i.e.,

(1) The sample mean is unbiased for the population mean.

(2) The sample variance $s_n^2 = \frac{1}{n-1}\sum_{k=1}^n (X_k - \bar{X}_n)^2$ is unbiased for the population variance. But $V_n = \frac{1}{n}\sum_{k=1}^n (X_k - \bar{X}_n)^2$ is not, in fact $\mathbf{E}[V_n] = \frac{n-1}{n}\sigma^2$.

It appears that s_n^2 is better, but the following remark says that one should be cautious in making such a statement.

Remark 12. In case of $N(\mu, \sigma^2)$ data, it turns out that although s_n^2 is unbiased and V_n is biased, the mean squared error of V_n is smaller! Further V_n is the maximum likelihood estimate of σ^2! Overall, unbiasedness is not so important as having smaller mean squared error, but for estimating variance (when the mean is not known), we always use s_n^2. The computation of the m.s.e is a bit tedious, so we skip it here.

Example 13. Let $X_1, \ldots, X_n$ be i.i.d. Ber(p). Then $\bar{X}_n$ is an estimate for p. It is unbiased since $\mathbf{E}[\bar{X}_n] = p$. Hence, the m.s.e of $\bar{X}_n$ is just the variance which is equal to $p(1-p)/n$.

A puzzle: A coin C_1 has probability p of turning up head and a coin C_2 has probability $2p$ of turning up head. All we know is that $0 < p < \frac{1}{2}$. You are given 20 tosses. You can choose all tosses from C_1 or all tosses from C_2 or some tosses from each (the total is 20). If the objective is to estimate p, what do you do?

Solution: If we choose to have all $n = 20$ tosses from C_1, then we get $X_1, \ldots, X_n$ that are i.i.d. Ber(p). An estimate for p is $\bar{X}_n$ which is unbiased and hence $\text{MSE}_{\bar{X}_n}(p) = \text{Var}(\bar{X}_n) = p(1-p)/n$. On the other hand if we choose to have all 20 tosses from C_2, then we get $Y_1, \ldots, Y_n$ that are i.i.d. Ber$(2p)$. The estimate for p is now $\bar{Y}_n/2$ which is also unbiased and has $\text{MSE}_{\bar{Y}_n/2}(p) = \text{Var}(\bar{Y}_n) = 2p(1-2p)/4 = p(1-2p)/2$. It is not hard to see that for all $p < 1/2$, $\text{MSE}_{\bar{Y}_n/2}(p) < \text{MSE}_{\bar{X}_n}(p)$ and hence choosing C_2 is better, at least by mean-squared criterion! It can be checked that if we choose to have k tosses from C_1 and the rest from C_2, the MSE of the corresponding estimate will be between the two MSEs found above and hence not better than $\bar{Y}_n/2$.

Another puzzle: A factory produces light bulbs having an exponential distribution with mean μ. Another factory produces light bulbs having an exponential distribution with mean 2μ. Your goal is to estimate μ. You are allowed to choose a total of 50 light bulbs (all from the first or all from the second or some from each factory). What do you do?

Solution: If we pick all $n = 50$ bulbs from the first factory, we see $X_1, \ldots, X_n$ i.i.d. Exp$(1/\mu)$. The estimate for μ is $\bar{X}_n$ which has $\text{MSE}_{\bar{X}_n}(\mu) = \text{Var}(\bar{X}_n) = \mu^2/n$. If we choose all bulbs from factory 2 we get $Y_1, \ldots, Y_n$ i.i.d. Exp$(1/2\mu)$. The estimate for μ is $\bar{Y}_n/2$. But $\text{MSE}_{\bar{Y}_n/2}(\mu) = \text{Var}(\bar{Y}_n/2) = (2\mu)^2/4n = \mu^2/n$. The two mean-squared errors are exactly the same!

Probabilistic thinking: Is there any calculation-free explanation why the answers to the two puzzles are as above? Yes, and it is illustrative of what may be called probabilistic thinking. Take the second puzzle. Why are the two estimates same by mean-squared error? Is one better by some other criterion?

Recall that if $X \sim \text{Exp}(1/\mu)$ then $X/2 \sim \text{Exp}(1/2\mu)$ and vice versa. Therefore, if we have data from Exp$(1/\mu)$ distribution, then we can divided all the numbers by 2 and convert it into data from Exp$(1/2\mu)$ distribution. Conversely if we have data from Exp$(1/2\mu)$ distribution, then we can convert it into data from Exp$(1/\mu)$ distribution by multiplying each number by 2. Hence there should be no advantage in choosing either factory. We leave it for you to think in analogous ways why in the first puzzle C_2 is better than C_1.

4. Confidence intervals

So far, in estimating of an unknown parameter, we give a single number as our guess for the known parameter. It would be better to give an interval and say with what confidence we expect the true parameter to lie within it. As a very simple example, suppose we have one random variable X with $N(\mu, 1)$ distribution. How do we estimate μ? Suppose the observed value of X is 2.7. Going by any method, the guess for μ would be 2.7 itself. But of course μ is not equal to X, so we would like to give an interval in which μ lies. How about $[X-1, X+1]$? Or $[X-2, X+2]$? Using normal tables, we see that $\mathbf{P}(X-1 < \mu < X+1) = \mathbf{P}(-1 < (X-\mu) < 1) = \mathbf{P}(-1 < Z < 1) \approx 0.68$ and similarly $\mathbf{P}(X - 2 < \mu < X + 2) \approx 0.95$. Thus, by making the interval longer we can be more confident that the true parameter lies within. But the accuracy of our statement goes down (if you want to know the average height of people in India, and the answer you give is "between 100cm and 200cm", it is very probably correct, but of little use!). The probability with which our CI contains the unknown parameter is called the level of confidence. Usually we fix the level of confidence, say as 0.90 and find an interval *as short as possible* but subject to the condition that it should have a confidence level of 0.90.

In this section we consider the problem of confidence intervals in Normal population. In the next we see a few other examples.

The setting: Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ random variables. We consider four situations.

(1) Confidence interval for μ when σ^2 is known.

(2) Confidence interval for σ^2 when μ is known.

(3) Confidence interval for μ when σ^2 is unknown.

(4) Confidence interval for σ^2 when μ is unknown.

A starting point in finding a confidence interval for a parameter is to first start with an estimate for the parameter. For example, in finding a CI for μ, we may start with $\bar{X}_n$ and enlarge it to an interval $[\bar{X}_n - a, \bar{X}_n + a]$. Similarly, in finding a CI for σ^2 we use the estimate $s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ if μ is unknown and $W_n = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2$ if the value of μ is known.

4.1. **Estimating μ when σ^2 is known.** We look for a confidence interval of the form $I_n = [\bar{X}_n - a, \bar{X}_n + a]$. Then,

$$\mathbf{P}\left(I_n \ni \mu\right) = \mathbf{P}\left(-a \le \bar{X}_n - \mu \le a\right) = \mathbf{P}\left(-\frac{a\sqrt{n}}{\sigma} \le \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \le \frac{a\sqrt{n}}{\sigma}\right)$$

Now we use two facts about normal distribution that we have seen before.

(1) If $Y \sim N(\mu, \sigma^2)$ then $aX + b \sim N(a\mu + b, a^2\sigma^2)$.

(2) If $Y_1 \sim N(\mu, \sigma^2)$ and $Y_2 \sim N(\nu, \tau^2)$ and they are independent, then $X + Y \sim N(\mu + \nu, \sigma^2 + \tau^2)$.

Consequently, $\bar{X}_n \sim N(0, \sigma^2/n)$ and $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim N(0, 1)$. Therefore,

$$\mathbf{P}\left(I_n \ni \mu\right) = \mathbf{P}(-\frac{a\sqrt{n}}{\sigma} \le Z \le -\frac{a\sqrt{n}}{\sigma})$$

where $Z \sim N(0, 1)$. Fix any $0 < \alpha < 1$ and denote by z_α the number such that $\mathbf{P}(Z > z_\alpha) = \alpha$ (in other words, z_α is the $(1-\alpha)$-quantile of the standard normal distribution). For example, from normal tables we find that $z_{0.05} \approx 1.65$ and $z_{0.005} \approx 2.58$ etc.

If we set $a = z_{\alpha/2}\sigma/\sqrt{n}$, we get

$$\mathbf{P}\left(\left[\bar{X}_n - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \bar{X}_n + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right] \ni \mu\right) = 1 - \alpha.$$

This is our confidence interval.

4.2. **Estimating σ^2 when μ is known.** Since μ is known, we use $W_n = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2$ to estimate σ^2. Here is an exercise.

Exercise 14. Let $Z_1, \ldots, Z_n$ be i.i.d. $N(0, 1)$ random variables. Then, $Z_1^2 + \ldots + Z_n^2 \sim \text{Gamma}(n/2, 1/2)$.

Solution: For $t > 0$ we have

$$\mathbf{P}\{Z_1^2 \leq t\} = \mathbf{P}\{-\sqrt{t} \leq Z_1 \leq \sqrt{t}\} = 2\int_0^{\sqrt{t}} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du = \frac{1}{\sqrt{2\pi}} \int_0^t e^{-s/2} s^{-1/2} ds.$$

Differentiate w.r.t t to see that the density of Z_1^2 is $h(t) = \frac{1}{\sqrt{\pi}} e^{-t/2} t^{-1/2} \sqrt{(1/2)}$, which is just the Gamma$(\frac{1}{2}, \frac{1}{2})$ density.

Now, each Z_k^2 has the same Gamma$(\frac{1}{2}, \frac{1}{2})$ density, and they are independent. Earlier we have seen that when we add independent Gamma random variables with the same scale parameter, the sum has a Gamma distribution with the same scale but whose shape parameter is the sum of the shape parameters of the individual summands. Therefore, $Z_1^2 + \ldots + Z_n^2$ has Gamma$(n/2, 1/2)$ distribution. This completes the solution to the exercise.

In statistics, the distribution Gamma$(1/2, 1/2)$ is usually called the *chi-squared distribution with n degrees of freedom*. Let $\chi_n^2(\alpha)$ denote the $1-\alpha$ quantile of this distribution. Similarly, $\chi_n^2(1-\alpha)$ is the α quantile (i.e., the probability for the chi-squared random variable to fall below $\chi_n^2(1-\alpha)$ is exactly α).

When X_i are i.i.d. $N(\mu, \sigma^2)$, we know that $(X_i - \mu)/\sigma$ are i.i.d. $N(0,1)$. Hence, by the above fact, we see that

$$\frac{nW_n}{\sigma^2} = \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2$$

has chi-squared distribution with n degrees of freedom. Hence

$$\mathbf{P}\left\{\frac{nW_n}{\chi_n^2\left(\frac{\alpha}{2}\right)} \leq \sigma^2 \leq \frac{nW_n}{\chi_n^2\left(1-\frac{\alpha}{2}\right)}\right\} = \mathbf{P}\left\{\chi_n^2\left(1-\frac{\alpha}{2}\right) \leq \frac{nW_n}{\sigma^2} \leq \chi_n^2\left(\frac{\alpha}{2}\right)\right\} = 1-\alpha.$$

Thus, $\left[\frac{ns_n^2}{\chi_{n-1}^2\left(\frac{\alpha}{2}\right)}, \frac{ns_n^2}{\chi_{n-1}^2\left(1-\frac{\alpha}{2}\right)}\right]$ is a $(1-\alpha)$-confidence interval for σ^2.

An important result: Before going to the next two confidence interval problems, let us try to understand the two examples already covered. In both cases, we came up with a random variable ($\sqrt{n}(\bar{X}_n - \mu)/\sigma$ and W_n/σ^2, respectively) which involved the data and the unknown parameter whose distributions we knew (standard normal and χ_n^2, respectively) and these distributions do not depend on any parameters. This is generally the key step in any confidence interval problem. For the next two problems, we cannot use the same two random variables as above as they depend on the other unknown parameter too (i.e., $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ uses σ which will be unknown and W_n/σ^2 uses μ which will be unknown). Hence, we need a new result that we state without proof.

Theorem 15. *Let $Z_1, \ldots, Z_n$ be i.i.d. $N(\mu, \sigma^2)$ random variables. Let $\bar{Z}_n$ and s_n^2 be the sample mean and the sample variance, respectively. Then,*

$$\bar{Z}_n \sim N(\mu, \frac{\sigma^2}{n}), \quad \frac{(n-1)s_n^2}{\sigma^2} \sim \chi^2_{n-1},$$

and the two are independent.

This is not too hard to prove (a muscle-flexing exercise in change of variable formula) but we skip the proof. Note two important features. First, the surprising independence of the sample mean and the sample variance. Second, the sample variance (appropriately scaled) has χ^2 distribution, just like W_n in the previous example, but the degree of freedom is reduced by 1. Now we use this theorem in computing confidence intervals.

4.3. **Estimating σ^2 when μ is unknown.** The estimate s_n^2 must be used as W_n depends on μ which is unknown. Theorem thm:indepofsamplemeanandvar tells us that $\frac{(n-1)s_n^2}{\sigma^2} \sim \chi^2_{n-1}$. Hence, by the same logic as before we get

$$\mathbf{P}\left\{\frac{(n-1)s_n^2}{\chi^2_{n-1}\left(\frac{\alpha}{2}\right)} \leq \sigma^2 \leq \frac{(n-1)s_n^2}{\chi^2_{n-1}\left(1-\frac{\alpha}{2}\right)}\right\} = \mathbf{P}\left\{\chi^2_{n-1}\left(1-\frac{\alpha}{2}\right) \leq \frac{(n-1)s_n^2}{\sigma^2} \leq \chi^2_{n-1}\left(\frac{\alpha}{2}\right)\right\}$$

$$= 1-\alpha.$$

Thus, $\left[\frac{(n-1)s_n^2}{\chi^2_{n-1}\left(\frac{\alpha}{2}\right)}, \frac{(n-1)s_n^2}{\chi^2_{n-1}\left(1-\frac{\alpha}{2}\right)}\right]$ is a $(1-\alpha)$-confidence interval for σ^2.

If μ is known, we could use the earlier confidence interval using W_n, or simply ignore the knowledge of μ and use the above confidence interval using s_n^2. What is the difference? The cost of ignoring the knowledge of μ is that the second confidence interval will be typically larger, although for large n the difference is slight. On the other hand, if our knowledge of μ was inaccurate, then the first confidence interval is invalid (we have no idea what its level of confidence is!) which is more serious. In realistic situations it is unlikely that we will know one of the parameters but not the other - hence, most often one just uses the confidence interval based on s_n^2.

4.4. **Estimating μ when σ^2 is unknown.** The earlier confidence interval We look for a confidence interval $[\bar{X}_n - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \bar{X}_n + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}]$ cannot be used as we do not know the value of σ.

A natural idea would be to use the estimate $s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ in place of σ^2. However, recall that the earlier confidence interval (in particular, the cut-off values $z_{\alpha/2}$ in the CI) was an outcome of the fact that

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim N(0,1).$$

Is it true if σ is replaced by s_n? Actually no, but we have a different distribution called *Student's t-distribution*.

Exercise 16. Let $Z \sim N(0,1)$ and $S^2 \sim \chi^2_n$ be independent. Then, the density of $\frac{Z}{S/\sqrt{n}}$ is given by

$$\frac{1}{\sqrt{n-1}\text{Beta}(\frac{1}{2}, \frac{n-1}{2})} \frac{1}{\left(1+\frac{t^2}{n-1}\right)^{\frac{n}{2}}}$$

for all $t \in \mathbb{R}$. This is known as *Student's t-distribution*.

The exact density of t-distribution is not important to remember, so the above exercise is optional. The point is that it can be computed from the change of variable formula and that by numerical integration its CDF can be tabulated.

How does this help us? From Theorem 15 we know that $\frac{\sqrt{n}(\bar{X}_n-\mu)}{\sigma} \sim N(0,1)$, $\frac{(n-1)s_n^2}{\sigma^2} \sim \chi^2_{n-1}$, and the two are independent. Take these random variables in the above exercise to conclude that $\frac{\sqrt{n}(\bar{X}_n-\mu)}{s_n}$ has t_{n-1} distribution.

The t-distribution is symmetric about zero (the density at t and at $-t$ are the same). Further, as the number of degrees of freedom goes to infinity, the t-density converges to the standard normal density. What we need to know is that there are tables from which we can read off specific quantiles of the distribution. In particular, by $t_n(\alpha)$ we mean the $1-\alpha$ quantile of the t-distribution with n degrees of freedom. Then of course, the α quantile is $-t_n(\alpha)$.

Returning to the problem of the confidence interval, from the fact stated above, we see that (use T_n to indicate a random variable having t-distribution with n degrees of freedom).

$$\begin{aligned}
&\mathbf{P}\left(\bar{X}_n - \frac{s_n}{\sqrt{n}} t_{n-1}\left(\frac{\alpha}{2}\right) \leq \mu \leq \bar{X}_n + \frac{s_n}{\sqrt{n}} t_{n-1}\left(\frac{\alpha}{2}\right)\right) \\
&= \mathbf{P}\left(-t_{n-1}\left(\frac{\alpha}{2}\right) \leq \frac{\sqrt{n}(\bar{X}_n-\mu)}{s_n} \leq t_{n-1}\left(\frac{\alpha}{2}\right)\right) \\
&= \mathbf{P}\left(-t_{n-1}\left(\frac{\alpha}{2}\right) \leq T_{n-1} \leq t_{n-1}\left(\frac{\alpha}{2}\right)\right) \\
&= 1-\alpha.
\end{aligned}$$

Hence, our $(1-\alpha)$-confidence interval is $\left[\bar{X}_n - \frac{s_n}{\sqrt{n}} t_{n-1}\left(\frac{\alpha}{2}\right), \bar{X}_n + \frac{s_n}{\sqrt{n}} t_{n-1}\left(\frac{\alpha}{2}\right)\right]$.

Remark 17. We remarked earlier that as $n \to \infty$, the t_{n-1} density approaches the standard normal density. Hence, $t_{n-1}(\alpha)$ approaches z_α for any α (this can be seen by looking at the t-table for large degree of freedom). Therefore, when n is large, we may as well use

$$\left[\bar{X}_n - \frac{s_n}{\sqrt{n}} z_{\alpha/2}, \bar{X}_n + \frac{s_n}{\sqrt{n}} z_{\alpha/2}\right].$$

Strictly speaking the level of confidence is smaller than for the one with $t_{n-1}(\alpha/2)$. However for n large the level of confidence is quite close to $1-\alpha$.

5. Remarks on the probabilty calculations involved in the previous section

There were a few facts used in this section. We summarize them here. Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$. Then

(1) $\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$.

(2) $\frac{nW_n}{\sigma^2} \sim \chi^2_n$ where $W_n = \frac{1}{n}\sum_{k=1}^{n-1}(X_k - \mu)^2$.

(3) $\frac{(n-1)s_n^2}{\sigma^2} \sim \chi^2_{n-1}$ where $s_n^2 == \frac{1}{n-1}\sum_{k=1}^{n-1}(X_k - \bar{X}_n)^2$.

(4) $\bar{X}_n$ and s_n^2 are independent.

We have used the first one many times. The second is also familiar, since $Z_i = (X_i - \mu)/\sigma$ are i.i.d. $N(0,1)$ variables and we have seen that sum of squares of n i.i.d. $N(0,1)$ variables has $|chi^2_n$ distribution. It remains to show the last two facts. They can be done together as follows.

Firstly, we use the standardized variables $Z_i = (X_i - \mu)/\sigma$ which are i.i.d. $N(0,1)$. The goal is to show that $\bar{Z}_n$ and $\sum_{k=1}^n (Z_k - \bar{Z}_n)^2$ are independent and the latter has χ^2_{n-1} distribution. Define

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,n} \\ a_{2,1} & a_{2,2} & \ldots & a_{2,n} \\ \vdots & \vdots & \ldots & \vdots \\ a_{n,1} & a_{n,2} & \ldots & a_{n,n} \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{bmatrix}$$

where the matrix is chosen as follows: Let the entries in the first row be $1/\sqrt{n}$. Then the first row is a unit vector. It can be extended to an orthonormal basis of $\mathbb{R}^n$. The other vectors in this basis will be the second, third,... rows of the matrix. There is a lot of choice, but it does not matter how we pick the orthonormal basis. With this, the matrix $A = (a_{i,j})_{i,j \le n}$ becomes an orthogonal matrix, i.e., $AA^t = I$. Because of this, $Y_1, \ldots, Y_n$ are also i.i.d. $N(0,1)$ random variables (check!).

Further, $Y_1 = \sqrt{n}Z_1$. From the orthogonality of A, it follows that $Y_1^2 + \ldots + Y_n^2 = Z_1^2 + \ldots + Z_n^2$ (because $\|AZ\|^2 = Z^t A^t A Z = Z^t Z$). Consequently,

$$\begin{aligned} Y_2^2 + \ldots + Y_n^2 &= Z_1^2 + \ldots + Z_n^2 - n\bar{Z}_n^2 \\ &= (Z_1 - \bar{Z}_n)^2 + \ldots + (Z_n - \bar{Z}_n)^2. \end{aligned}$$

This shows that $(Z_1 - \bar{Z}_n)^2 + \ldots + (Z_n - \bar{Z}_n)^2$ depends only on $Y_2, \ldots, Y_n$ and hence is independent of $\bar{Z}_n$ which depends on Y_1 alone. Further, $Y_2^2 + \ldots + Y_n^2$ has χ^2_{n-1} distribution, being a sum of squares of $(n-1)$ i.i.d. $N(0,1)$ variables. This completes the proof. ■

6. Confidence interval for the mean

Now suppose $X_1, \ldots, X_n$ are i.i.d. random variables from some distribution with mean μ and variance σ^2, both unknown. How can we construct a confidence interval for μ?

In case of normal distribution, recall that the $(1-\alpha)$-CI that we gave was

$$\left[\bar{X}_n - \frac{s_n}{\sqrt{n}} t_{n-1}\left(\frac{\alpha}{2}\right), \bar{X}_n + \frac{s_n}{\sqrt{n}} t_{n-1}\left(\frac{\alpha}{2}\right)\right] \text{ or } \left[\bar{X}_n - \frac{s_n}{\sqrt{n}} z_{\alpha/2}, \bar{X}_n + \frac{s_n}{\sqrt{n}} z_{\alpha/2}\right]$$

Is this a valid confidence interval in general? The answer is "No" for both. If X_i are from some general distribution then the distributions of $\sqrt{n}(\bar{X}_n - \mu)/s_n$ and $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ are very complicated to find. Even if X_i come from binomial or exponential family, these distributions will depend on the parameters in a complex way (in particular, the distributions are not free from the parameters, which is important in constructing confidence intervals).

But suppose n is large. Then the sample variance is close to population variance and hence $s_n \approx \sigma$. Further, by CLT, we know that $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ has approximately $N(0,1)$ distribution. Hence, we see that

$$\mathbf{P}\left\{-z_{\alpha/2} \le \frac{\sqrt{n}(\bar{X}_n - \mu)}{s_n} \le z_{\alpha/2}\right\} \approx \Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2}) = 1 - \alpha.$$

Consequently, we may say that

$$\mathbf{P}\left\{\bar{X}_n - \frac{s_n}{\sqrt{n}} z_{\alpha/2} \le \mu \le \bar{X}_n + \frac{s_n}{\sqrt{n}} z_{\alpha/2}\right\} \approx 1 - \alpha.$$

Thus, $\left[\bar{X}_n - \frac{s_n}{\sqrt{n}} z_{\alpha/2}, \bar{X}_n + \frac{s_n}{\sqrt{n}} z_{\alpha/2}\right]$ is an approximate $(1-\alpha)$-confidence interval. Further, when n is large, the difference between $V_n = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ and $V_n := \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ is small (indeed, $s_n^2 = (n/(n-1))V_n$). Hence it is also okay to use $\left[\bar{X}_n - \frac{\sqrt{V_n}}{\sqrt{n}} z_{\alpha/2}, \bar{X}_n + \frac{\sqrt{V_n}}{\sqrt{n}} z_{\alpha/2}\right]$ as an approximate $(1-\alpha)$-confidence interval.

Example 18. Let $X_1, \ldots, X_n$ be i.i.d. Ber(p). Consider the problem of finding a confidence interval for p. Since each X_i is 0 or 1, observe that

$$\hat{s}_n^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}_n^2 = \bar{X}_n - (\bar{X}_n)^2 = \bar{X}_n(1 - \bar{X}_n).$$

Hence, an approximate $(1-\alpha)$-CI for p is given by

$$\left[\bar{X}_n - z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}, \bar{X}_n + z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right].$$

7. A DIGRESSION - THE BAYESIAN FRAMEWORK

For the sake of coherence, we have given one framework of statistics, but there are others. In our framework, the unknown is not equated with random. Like the laws of physics, some of which are unknown (and those known today were unknown some centuries ago) but not random, we treat the parameters as unknown but fixed. In the *Beysian framework*, one treats the unknown

parameters as random. We shall give a brief introduction to this, in the context of a problem of confidence interval.

7.1. **Beyesian approach to estimating the success probability of a coin.** Let $X_1, \ldots, X_n$ be tosses from a coin of unknown probability p of success. If we have no inkling of where p is, we assume that it is anywhere in $[0,1]$, distributed uniformly at random. To write it out explicitly, the assumptions are

(1) $p \sim \text{unif}[0,1]$. This unif$[0,1]$ is called the *prior distribution* of p.

(2) Conditional on p, the tosses $X_1, \ldots, X_n$ are i.i.d. Ber(p).

It may be noted that in this setting X_is are not independent unconditionally (if 9 out of 10 tosses are heads, we guess that perhaps the p had come out high, hence our guess is that the next toss is very likely to be a head).

Then, the conditional distribution of p given the data $X_1, \ldots, X_n$ is computed. This is basically Bayes rule, except that there are uncountably many possible values of p (the different values of p are like $A_1, A_2, \ldots$ and the data is like the event B, then we compute $\mathbf{P}(A_i \,\big|\, B)$ using Bayes' rule as the ratio of $\mathbf{P}(A_i)\mathbf{P}(B \,\big|\, A_i)$ to the sum of such quantities over i). In our setting this gives the conditional density of p given $X_1, \ldots, X_n$ as

$$\frac{\binom{n}{k} p^k (1-p)^{n-k}}{\int_0^1 \binom{n}{k} x^k (1-x)^{n-k} dx}$$

where $k = X_1 + \ldots + X_n$ is the number of heads. This is called the *posterior density* of p. Writing $k = n\bar{X}_n$, we see that the posterior distribution is just Beta$(1 + n\bar{X}_n, 1 + n(1-\bar{X}_n))$.

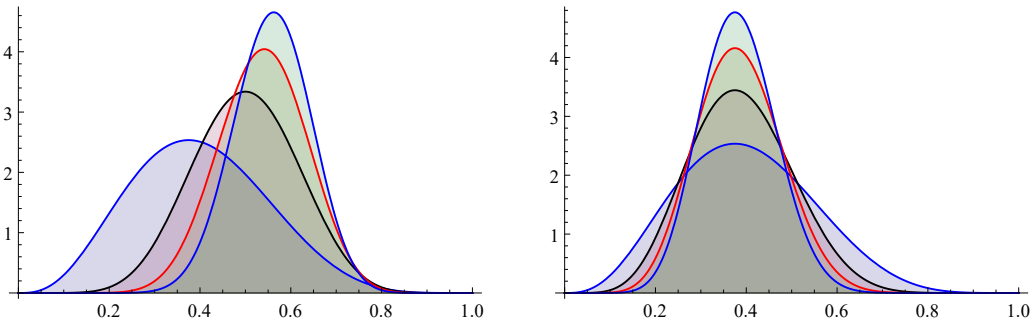

FIGURE 1. Tosses are made a fair coin. The posterior distributions (assuming uniform prior) for p with $n = 8, 16, 24, 32$ are plotted. The narrower distributions correspond to higher n. The second figure is the same, for a different set of data.

How is this useful? If our goal is to give an estimate for p, we just give the expected value of the posterior distribution. In the case at hand (since the expected value of Beta(p,q) distribution is $p/(p+q)$) this give $\hat{p} = \frac{1+n\bar{X}_n}{n+1}$. This is almost the same as $\bar{X}_n$ for large n, but for small n, the

prior assumption that p could be anywhere makes it a bit different. In general, the estimates in this framework depend on the data and also on the prior. As the data size increases, the effect of the prior fades away.

If we need a confidence interval for p, we find numbers $a < b$ such that the posterior distribution puts mass $\alpha/2$ below a and mass $\alpha/2$ above b. Then declare $[a, b]$ to be the confidence interval. In the case at hand, we shall require the quantiles of the Beta distributions, which can be found on a computer. With $\alpha = 0.1$ (and giving up $\alpha/2$ equally on both sides), the confidence intervals for the four graphs shown in Figure 7.1 turn out to be $[0.16875, 0.655059]$, $[0.31083, 0.68917]$, $[0.378622, 0.69487]$ and $[0.418562, 0.69509]$. The confidence intervals become narrower with larger samples.

8. Actual confidence by simulation

Suppose we have a candidate confidence interval whose confidence we do not know. For example, let us take the confidence interval

$$\left[\bar{X}_n - z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}, \bar{X}_n + z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right].$$

for the parameter p of i.i.d. Ber(p) samples. We saw that for large n this has approximately $(1-\alpha)$ confidence. But how large is large? One way to check this is by simulation. We explain how.

Take $p = 0.3$ and $n = 10$. Simulate $n = 10$ independent Ber(p) random variables and compute the confidence interval given above. Check whether it contains the true value of p (i.e., 0.3) or not. Repeat this exercise 10000 times and see what proportion of times it contains 0.3. That proportion is the true confidence, as opposed to $1-\alpha$ (which is valid only for large n). Repeat this experiment with $n = 20$, $n = 30$ etc. See how close the actual confidence is to $1 - \alpha$. Repeat this experiment with different value of p. The n you need to get close to $1 - \alpha$ will depend on p (in particular, on how close p is to $1/2$).

This was about checking the validity of a confidence interval that was specified. In a real situation, it may be that we can only get $n = 20$ samples. Then what can we do? If we have an idea of the approximate value of p, we can first simulate Ber(p) random numbers on a computer. We compute the sample mean each time, and repeat 10000 times to get so many values of the sample mean. Note that the histogram of these 10000 values tells us (approximately) the actual distribution of $\bar{X}_n$. Then we can find t (numerically) such that $[\bar{X}_n - t, \bar{X}_n + t]$ contains the true value of p in $(1 - \alpha)$-proportion of the 10000 trials. Then, $[\bar{X}_n - t, \bar{X}_n + t]$ is a $(1 - \alpha)$-CI for p. Alternately, we may try a CI of the form

$$\left[\bar{X}_n - t\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}, \bar{X}_n + t\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right].$$

where we choose t numerically to get $(1 - \alpha)$ confidence.

Summary: The gist of this discussion is this. In the neatly worked out examples of the previous sections, we got explicit confidence intervals. But we assumed that we knew the data came from $N(\mu, \sigma^2)$ distribution. What if that is not quite right? What if it is not any of the nicely studied distributions? The results also become invalid in such cases. For large n, using law of large numbers and CLT we could overcome this issue. But for small n? The point is that using simulations we can calculate probabilities, distributions, etc, numerically and approximately. That is often better, since it is more robust to assumptions.

9. Hypothesis testing - first examples

Earlier in the course we discussed the problem of how to test whether a "psychic" can make predictions better than a random guesser. This is a prototype of what are called *testing problems*. We start with this simple example and introduce various general terms and notions in the context of this problem.

Question 19. A "psychic" claims to guess the order of cards in a deck. We shuffle a deck of cards, ask her to guess and count the number of correct guesses, say X.

One hypotheses (we call it the *null hypothesis* and denote it by H_0) is that the psychic is guessing randomly. The *alternate hypothesis* (denoted H_1) is that his/her guesses are better than random guessing (in itself this does not imply existence of psychic powers. It could be that he/she has managed to see some of the cards etc.). Can we decide between the two hypotheses based on X?

What we need is a rule for deciding which hypothesis is true. A rule for deciding between the hypotheses is called a *test*. For example, the following are examples of rules (the only condition is that the rule must depend only on the data at hand).

Example 20. We present three possible rules.

(1) If X is an even number declare that H_1 is true. Else declare that H_1 is false.

(2) If $X \geq 5$, then accept H_1, else reject H_1.

(3) If $X \geq 8$, then accept H_1, else reject H_1.

The first rule does not make much sense as the parity (evenness or oddness) has little to do with either hypothesis. On the other hand, the other two rules make some sense. They rely on the fact that if H_1 is true then we expect X to be larger than if H_0 is true. But the question still remains, should we draw the line at 5 or at 8 or somewhere else?

In testing problems there is only one objective, to avoid the following two possible types of mistakes.

Type-I error: H_0 is true but our rule concludes H_1.

Type-II error: H_1 is true but our rule concludes H_0.

The probability of type-I error is called the *significance level* of the test and usually denote by α. That is, $\alpha = \mathbf{P}_{H_0}\{\text{the test accepts } H_1\}$ where we write $\mathbf{P}_{H_0}$ to mean that the probability is calculated under the assumption that H_0 is true. Similarly one define the *power* of the test as $\beta = \mathbf{P}_{H_1}\{\text{the test accepts } H_1\}$. Note that β is the probability of not making type-II error, and hence we would like it to be close to 1. Given two tests with the same level of significance, the one with higher power is better. Ideally we would like both to be small, but that is not always achievable.

We fix the desired level of significance, usually $\alpha = 0.05$ or 0.1 and only consider tests whose probability of type-I error is at most α. It may seem surprising that we take α to be so small. Indeed the two hypotheses are not treated equally. Usually H_0 is the default option, representing traditional belief and H_1 is a claim that must prove itself. As such, the burden of proof is on H_1.

To use analogy with law, when a person is convicted, there are two hypotheses, one that he is guilty and the other that he is not guilty. According to the maxim "innocent till proved guilty", one is not required to prove his/her innocence. On the other hand guilt must be proved. Thus the null hypothesis is "not guilty" and the alternative hypothesis is "guilty".

In our example of card-guessing, assuming random guessing, we have calculated the distribution of X long ago. Let $p_k = \mathbf{P}\{X = k\}$ for $k = 0, 1, \ldots, 52$. Now consider a test of the form "Accept H_1 if $X \geq k_0$ and reject otherwise". Its level of significance is

$$\mathbf{P}_{H_0}\{\text{accept } H_1\} = \mathbf{P}_{H_0}\{X \geq k_0\} = \sum_{i=k_0}^{52} p_i.$$

For $k_0 = 0$, the right side is 1 while for $k_0 = 52$ it is $1/52!$ which is tiny. As we increase k_0 there is a first time where it becomes less than or equal to α. We take that k_0 to be the threshold for cut-off.

In the same example of card-guessing, let $\alpha = 0.01$. Let us also assume that Poisson approximation holds. This means that $p_j \approx e^{-1}/j!$ for each j. Then, we are looking for the smallest k_0 such that $\sum_{j=k_0}^{\infty} e^{-1}/j! \leq 0.01$. For $k_0 = 4$, this sum is about 0.019 while for $k_0 = 5$ this sum is 0.004. Hence, we take $k_0 = 5$. In other words, accept H_1 if $X \geq 5$ and reject if $X < 5$. If we took $\alpha = 0.0001$ we would get $k_0 = 7$ and so on.

Strength of evidence: Rather than merely say that we accepted H_1 or rejected it would be better to say how strong the evidence is in favour of the alternative hypothesis. This is captured by the *p-value*, a central concept of decision making. It is defined as *the probability that data drawn from the null hypothesis would show closer agreement with the alternative hypothesis than the data we have at hand* (read it five times!).

Before we compute it in our example, let us return to the analogy with law. Suppose a man is convicted for murder. Recall that H_0 is that he is not guilty and H_1 is that he is guilty. Suppose his fingerprints were found in the house of the murdered person. Does it prove his guilt? It is some evidence in favour of it, but not necessarily strong. For example, if the convict was a friend

of the murdered person, then he might be innocent but have left his fingerprints on his visits to his friend. However if the convict is a total stranger, then one wonders why, if he was innocent, his finger prints were found there. The evidence is stronger for guilt. If bloodstains are found on his shirt, the evidence would be even stronger! In saying this, we are asking ourselves questions like "if he was innocent, how likely is it that his shirt is blood-stained?". That is p-value. Smaller the p-value, stronger the evidence for the alternate hypothesis.

Now we return to our example. Suppose the observed value is $X_{\text{obs}} = 4$. Then the p-value is $\mathbf{P}\{X \geq 4\} = p_4 + \ldots + p_{52} \approx 0.019$. If the observed value was $X_{\text{obs}} = 6$, then the p-value would be $p_6 + \ldots + p_{52} \approx 0.00059$. Note that the computation of p-value does not depend on the level of significance. It just depends on the given hypotheses and the chosen test.

10. Testing for the mean of a normal population

Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$. We shall consider the following hypothesis testing problems.

(1) One sided test for the mean. $H_0 : \ \mu = \mu_0$ versus $H_1 : \ \mu > \mu_0$.

(2) Two sided test for the mean. $H_0 : \ \mu = \mu_0$ versus $H_1 : \ \mu \neq \mu_0$.

This kind of problem arises in many situations in comparing the effect of a treatment as follows.

Example 21. Consider a drug claimed to reduce blood pressure. How do we check if it actually does? We take a random sample of n patients, measure their blood pressures $Y_1, \ldots, Y_n$. We administer the drug to each of them and again measure the blood pressures $Y_1', \ldots, Y_n'$, respectively. Then, the question is whether the mean blood pressure decreases upon giving the treatment. To this effect, we define $X_i = Y_i - Y_i'$ and wish to test the hypothesis that the mean of X_is is strictly positive. If X_i are indeed normally distributed, this is exactly the one-sided test above.

Example 22. The same applies to test the efficacy of a fertilizer to increase yield, a proposed drug to decrease weight, a particular educational method to improve a skill, or a particular course such as the current *probability and statistics course* in increasing subject knowledge. To make a policy decision on such matters, we can conduct an experiment as in the above example.

For example, a bunch of students are tested on probability and statistics and their scores are noted. Then they are subjected to the course for a semester. They are tested again after the course (for the same marks, and at the same level of difficulty) and the scores are again noted. Take differences of the scores before and after, and test whether the mean of these differences is positive (or negative, depending on how you take the difference). This is a one-sided tests for the mean. Note that in these examples, we are taking the null hypothesis to be that there is no effect. In other words, the burden of proof is on the new drug or fertilizer or the instructor of the course.

The test: Now we present the test. We shall use the statistic $\mathcal{T} := \frac{\sqrt{n}(\bar{X}-\mu_0)}{s}$ where $\bar{X}$ and s are the sample mean and sample standard deviation.

(1) In the one-sided test, we accept the alternative hypothesis if $\mathcal{T} > t_{n-1}(\alpha)$.

(2) In the two sided-test, accept the alternative hypothesis if $\mathcal{T} > t_{n-1}(\alpha/2)$ or $\mathcal{T} < -t_{n-1}(\alpha/2)$.

The rationale behind the tests: If $\bar{X}$ is much larger than μ_0 then the greater is the evidence that the true mean μ is greater than μ_0. However, the magnitude depends on the standard deviation and hence we divide by s (if we knew σ we would divide by that). Another way to see that this is reasonable is that $\mathcal{T}$ does not depend on the units in which you measure X_is (whether X_i are measured in meters or centimeters, the value of $\mathcal{T}$ does not change).

The significance level is α: The question is where to draw the threshold. We have seen before that *under the null hypothesis* $\mathcal{T}$ has a t_{n-1} distribution. Recall that this is because, if the null hypothesis is true, then $\frac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma} \sim N(0,1)$, $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$ and the two are independent. Thus, the given tests have significance level α for the two problems.

Remark 23. Earlier we considered the problem of constructing a $(1-\alpha)$-CI for μ when σ^2 is unknown. The two sided test abovecan be simply stated as follows: Accept the alternative at level α if the corresponding $(1-\alpha)$-CI does not contain μ_0. Conversely, if we had dealt with testing problems first, we could define a confidence interval as the set of all those μ_0 for which the corresponding test rejects the alternative.

Thus, confidence intervals and testing are closely related. This is true in some greater generality. For example, we did not construct confidence interval for μ, but you should do so and check that it is closely related to the one-sided tests above.

11. Testing for the difference between means of two normal populations

Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu_1, \sigma_1^2)$ and let $Y_1, \ldots, Y_m$ be i.i.d. $N(\mu_2, \sigma_2^2)$. We shall consider the following hypothesis testing problems.

(1) One sided test for the difference in means. $H_0: \ \mu_1 = \mu_2$ versus $H_1: \ \mu_1 > \mu_2$.

(2) Two sided test for the mean. $H_0: \ \mu_1 = \mu_2$ versus $H_1: \ \mu_1 \neq \mu_2$.

This kind of problem arises in many situations in comparing two different populations or the effect of two different treatments etc. Actual data sets of such questions can be found in the homework.

Example 24. Suppose a new drug to reduce blood pressure is introduced by a pharmaceutical company. There is already an existing drug in the market which is working reasonably alright. But it is claimed by the company that the new drug is better. How to test this claim?

We take a random sample of $n+m$ patients and break them into two groups of n and of m patients. The first group is administered the new drug while the second group is administered the old drug. Let $X_1, \ldots, X_n$ be the *decrease in blood pressures* in the first group. Let $Y_1, \ldots, Y_m$ be the

decrease in blood pressures in the second group. The claim is that one average X_is are larger than Y_is.

Note that it does not make sense to subtract $X_i - Y_i$ and reduce to a one sample test as in the previous section (here X_i is a measurement on one person and Y_i is a measurement on a completely different person! Even the number of persons in the two groups may differ). This is an example of a two-sample test as formulated above.

Example 25. The same applies to many studies of comparision. If someone claims that Americans are taller than Indians on average, or if it is claimed that cycling a lot leads to increase in height, or if it is claimed that Chinese have higher IQ than Europeans, or if it is claimed that *Honda Activa* gives better mileage than *Suzuki Access*, etc., etc., the claims can be reduced to the two-sample testing problem as introduced above.

BIG ASSUMPTION: We shall assume that $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (yet unknown). This assumption is not made because it is natural or because it is often observed, but because it leads to mathematical simplification. Without this assumption, no exact level-α test has been found!

The test: Let $\bar{X}, \bar{Y}$ denote the sample means of X and Y and let s_X, s_Y denote the corresponding sample standard deviations. Since σ^2 is the assumed to be the same for both populations, s_X^2 and s_Y^2 can be combined to define

$$S^2 := \frac{(n-1)s_X^2 + (m-1)s_Y^2}{m+n-2}$$

which is a better estimate for σ^2 than just s_X^2 or s_Y^2 (this S^2 is better than simply taking $(s_X^2 + s_Y^2)/2$ because it gives greater weight to the larger sample).

Now define $\mathcal{T} = \sqrt{\frac{1}{n} + \frac{1}{m}} \left(\frac{\bar{X} - \bar{Y}}{S} \right)$. The following tests hav significance level α.

(1) For the one-sided test, accept the alternative if $\mathcal{T} > t_{n+m-2}(\alpha)$.

(2) For the one-sided test, accept the alternative if $\mathcal{T} > t_{n+m-2}(\alpha/2)$ or $\mathcal{T} < -t_{n+m-2}(\alpha/2)$.

The rationale behind the tests: If $\bar{X}$ is much larger than $\bar{Y}$ then the greater is the evidence that the true mean μ_1 is greater than μ_2. But again we need to standardize by dividing this by an estimate of σ, namely S. The resulting statistic $\mathcal{T}$ has a t_{m+n-2} distribution as explained below.

The significance level is α: The question is where to draw the threshold. From the facts we know,

$$\bar{X} \sim N(\mu_1, \sigma_1^2/n),$$

$$\bar{Y} \sim N(\mu_2, \sigma_2^2/m),$$

$$\frac{(n-1)}{\sigma^2} s_X^2 \sim \chi^2_{n-1},$$

$$\frac{(m-1)}{\sigma^2} s_Y^2 \sim \chi^2_{m-1}$$

and the four random variables are independent. From this, it follows that $(m+n-2)S^2$ has χ^2_{n+m-2} distribution. *Under the null hypothesis* $\frac{1}{\sigma}\sqrt{\frac{1}{n}+\frac{1}{m}}(\bar{X}-\bar{Y})$ has $N(0,1)$ distribution and is independent of S. Taking ratios, we see that $\mathcal{T}$ has t_{m+n-2} distribution (under the null hypothesis).

12. Testing for the mean in absence of normality

Suppose $X_1, \ldots, X_n$ are i.i.d. Ber(p). Consider the test

$$H_0 : \ p = p_0 \quad \text{versus} \quad H_1 : \ p \neq p_0.$$

One can also consider the one-sided test. Just as in the confidence interval problem, we can give a solution when n is large, using the approximation provided by the central limit theorem. Recall that an approximate $(1-\alpha)$-CI is

$$\left[\bar{X}_n - z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}, \bar{X}_n + z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right].$$

Inverting this confidence interval, we see that a reasonable test is:

Reject the alternative if p_0 belongs to the above CI. That is, accept the alternative if

$$\bar{X}_n - z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \leq p_0 \leq \bar{X}_n + z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}$$

This test has (approximately) significance level α.

More generally, if we have data $X_1, \ldots, X_n$ from a population with mean μ and variance σ^2, then consider the test

$$H_0 : \ \mu = \mu_0 \quad \text{versus} \quad H_1 : \ \mu \neq \mu_0.$$

A test with approximate significance level α is given by: Reject the alternative if

$$\bar{X}_n - z_{\alpha/2}\frac{s_n}{\sqrt{n}} \leq \mu_0 \leq \bar{X}_n + z_{\alpha/2}\frac{s_n}{\sqrt{n}}.$$

Just as with confidence intervals, we can find the actual level of significance (if n is not large enough) by simulating data on a computer.

13. CHI-SQUARED TEST FOR GOODNESS OF FIT

At various times we have made statements such as "heights follow normal distribution", "lifetimes of bulbs follow exponential distribution" etc. Where do such claims come from? Over years of analysing data, of course. This leads to an interesting question. Can we test whether lifetimes of bulbs do follow exponential distribution?

We start with a simple example of testing whether a die is fair. The hypotheses are H_0 : the die is fair, versus H_1 : the die is unfair[2].

We throw the die n times and record the observations $X_1, \ldots, X_n$. For $j \leq 6$, let O_j be the number of times we observe the face j turn up. In symbols $O_j = \sum_{i=1}^{n} \mathbf{1}_{X_i=j}$. Let $E_j = \mathbf{E}[O_j] = \frac{n}{6}$ be the expected number of times we see the face j (under the null hypothesis). Common sense says that if H_0 is true then O_j and E_j must be rather close for each j. How to measure the closeness? Karl Pearson introduced the test statistic

$$T := \sum_{j=1}^{6} \frac{(O_j - E_j)^2}{E_j}.$$

If the desired level of significance is α, then the Pearson χ^2-test says "Reject H_0 if $T \geq \chi^2_5(\alpha)$". The number of degrees of freedom is 5 here. In general, it is one less than the number of bins (i.e., how many terms you are summing to get T).

Some practical points: The χ^2 test is really an asymptotic statement. For large n, the level of significance is approximately $1 - \alpha$. There is no assurance for small n. Further, in performing the test, it is recommended that each bin must have at least 5 observations (i.e., $O_j \geq 5$). Otherwise we club together bins with fewer entries. The number 5 is a rule of thumb, the more the better.

Fitting the Poisson distribution: We consider the famous data collected by Rutherford, Chadwick and Ellis on the number of radioactive disintegrations. For details see the book of Feller's book (section VI.7) or this website.

The data consists of $X_1, \ldots, X_{2608}$ (where X_k is the number of particles detected by the counter in the k^{th} time interval. The hypotheses are

$$H_0 : \ F \text{ is a Poisson distribution.} \qquad H_1 : \ F \text{ is not Poisson.}$$

The physical theories predict that the distribution ought to be Poisson and hence we have taken it as the null hypothesis[3]

[2]You may feel that the null and alternative hypotheses are reversed. Is not independence a special property that should prove itself. Yes and no. Here we are imagining a situation where we have some reason to think that the die is fair. For example perhaps the die looks symmetric.

[3]When a new theory is proposed, it should prove itself and is put in the alterntive hypotheis, but here we take it as null.

We define O_j as the number of time intervals in which we see exactly j particles. Thus $O_j = \sum_{i=1}^{2608} \mathbf{1}_{X_i=j}$. How do we find the expected numbers? If the null hypothesis had said that F has Poisson(1) distribution, we could use that to find the expected numbers. But H_0 only says Poisson(λ) for an unspecified λ? This brings in a new feature.

First estimate λ, for example $\hat{\lambda} = \bar{X}_n$ is an MLE as well as method of moments estimate. Then we use this to calculate Poisson probabilities and the expected numbers. In other words, $E_j = e^{-\hat{\lambda}} \frac{\hat{\lambda}^j}{j!}$. For the given data we find that $\hat{\lambda} = 3.87$. The table is as follows.

j	0	1	2	3	4	5	6	7	8	9	≥ 10
O_j	57	203	383	525	532	408	273	139	45	27	16
E_j	54.4	210.5	407.4	525.4	508.4	393.5	253.8	140.3	67.9	29.2	17.1

Two remarks: The original data would have consisted of several more bins for $j = 11, 12 \ldots$. These have been clubbed together to perform the χ^2 test (instead of a minimum of 5 per bin, they may have ensured that there are at least 10 per bin). Also, the estimate $\hat{\lambda} = 3.87$ was obtained before clubbing these bins. Indeed, if the data is merely presented as the above table, there will be some ambiguity in how to find $\hat{\lambda}$ as one of the bins says "≥ 10".

Then we compute

$$T = \sum_{j=0}^{10} \frac{(O_j - E_j)^2}{E_j} = 14.7.$$

Where should we look up in the χ^2 table? Earlier we said that the degrees of freedom is one less than the number of bins. Here we give the more general rule.

$$\text{Degrees of freedom of the } \chi^2 = \text{ No. of bins } - 1 - \text{No. of parameters estimated from data.}$$

In our case we estimated one parameter, λ hence the d.f. of the χ^2 is $11 - 1 - 1 = 9$. Looking at χ^2_9 table one can see that the p-value is 0.10. This is the probability that a χ^2_9 random variable is greater than 14.7. (Caution: Elsewhere I see that the p-value for this experiment is reported as 0.17, please check my calculations!). This means that at 5% level, we would not reject the null hypothesis. If the p-value was 0.17, we would not reject the null hypothesis even at 10% level.

Fitting a continuous distribution: Chi-squared test can be used to test goodness of fit for continuous distributions too. We need some modifications. We must make bins of appropriate size, like $[a, a+h], [a+h, a+2h], \ldots, [a+h(k-1), a+hk]$ for a suitable h and k. Then we find the expected numbers in each bin using the null hypothesis (first estimating some parameters if necessary) and then proceed to compute T in the same way as before. Then check against the χ^2 table with the appropriate degrees of freedom. We omit details.

The probability theorem behind the χ^2-test for goodness of fit: Let $(W_1, \ldots, W_k)$ have multinomial distribution with parameters $n, m, (p_1, \ldots, p_k)$. (In other words, place n balls at random into m bins, but each ball goes into the i^{th} bin with probability p_i and distinct balls are assigned independently of each other). The following proposition is the mathematics behind Pearson's test.

Proposition [Pearson]: Fix $k, p_1, \ldots, p_k$. Let $T_n = \sum_{i=1}^{k} \frac{(W_i - np_i)^2}{np_i}$. Then T_n converges to a χ^2_{k-1} distribution in the sense that $\mathbf{P}\{T_n \leq x\} \to \int_0^x f_{k-1}(u)du$ where f_{k-1} is the density of χ^2_{k-1} distribution.

How does this help? Suppose $X_1, \ldots, X_n$ are i.i.d. random variables taking k values (does not matter what the values are, say $t_1, t_2, \ldots, t_k$) with probabilities $p_1, \ldots, p_k$. Then, let W_i be the number of X_is whose value is t_i. Clearly, $(W_1, \ldots, W_k)$ has a multinomial distribution. Therefore, for large n, the random variable T_n defined above (which is in fact the χ^2-statistic of Pearson) has approximately χ^2_{k-1} distribution. This explains the test.

Sketch of proof of the proposition: Start with the case $k = 2$. Then, $W_1 \sim \text{Bin}(n, p_1)$ and $W_2 = r - W_1$. Thus, $T_n = \frac{(W_1 - np_1)^2}{np_1p_2}$ (recall that $p_1 + p_2 = 1$ and check this!). We know that $(W_1 - np_1)/\sqrt{np_1q_1}$ is approximately a $N(0, 1)$ random variable, where $q_i = 1 - p_i$). Its square has (approximatelyχ^2_1 distribution. Thus the proposition is proved for $k = 2$.

When $k > 2$, what happens is that the random variables $\xi_i := (W_i - np_i)/\sqrt{np_iq_i}$ are approximately $N(0, 1)$, but not independent. In fact the correlation between ξ_i and ξ_j is close to $-\sqrt{p_ip_j/q_iq_j}$. The sum of squares of ξ_is gives the χ^2 statistic. On the other hand, one can (with some clever linear algebra/matrix manipulation) write $\sum_{i=1}^{k} \xi_i^2$ as $\sum_{i=1}^{k-1} \eta_i^2$ where η_i are *independent* $N(0, 1)$ random variables. Thus we get χ^2_{k-1} distribution.

14. Tests for independence

Suppose we have a bivariate sample $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ i.i.d. from a joint density (or joint pmf) $f(x, y)$. The question is to decide whether X_i is independent of Y_i.

Example 26. There are many situations in which such a problem arises. For example, suppose a bunch of students are given two exams, one testing mathematical skills and another testing verbal skills. The underlying goal may be to investigate whether the human brain has distinct centers for verbal and quantitative thinking.

Example 27. As another example, say we want to investigate whether smoking causes lung cancer. In this case, for each person in the sample, we take two measurements - X (equals 1 if smoker and 0 if not) and Y (equal 1 if the person has lung cancer, 0 if not). The resulting data may be

summarized in a two-way table as follows.

	$X=0$	$X=1$	
$Y=0$	$n_{0,0}$	$n_{0,1}$	$n_{0\cdot}$
$Y=1$	$n_{1,0}$	$n_{1,1}$	$n_{1\cdot}$
	$n_{\cdot 0}$	$n_{\cdot 1}$	n

Here the total sample is of n persons and $n_{i,j}$ denote the numbers in each of the four boxes. The numbers $n_{0\cdot}$ etc denote row or column sums. The statistical problem is to check if smoking (X) and incidence of lung cancer (Y) are positively correlated.

Testing independence in bivariate normal: We shall not discuss this problem in detail but instead quickly give some indicators and move on. Here we have (X_i, Y_i) i.i.d bivariate normal random variables with $\mathbf{E}[X] = \mu_1$, $\mathbf{E}[Y] = \mu_2$, $\text{Var}(X) = \sigma_1^2$, $\text{Var}(Y) = \sigma_2^2$ and $\text{Corr}(X, Y) = \rho$. The testing problem is $H_0 : \ \rho = 0$ versus $H_1 : \ \rho \neq 0$. (Remember that if (X, Y) is bivariate normal, then X and Y are independent if and only if X and Y are uncorrelated.

The natural statistic to consider is the sample correlation coefficient (*Pearson's r statistic*)

$$r_n := \frac{s_{X,Y}}{s_X . s_Y}$$

where s_X^2, s_Y^2 are the sample variances of X and Y and $s_{X,Y} = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$ is the sample covariance. It is clear that the test should reject null hypothesis if r_n is away from 0. To decide the threshold we need the distribution of r_n under the null hypothesis.

Fisher: Under the null hypothesis, r_n^2 has Beta$(\frac{1}{2}, \frac{n-2}{2})$ distribution.

Using this result, we can draw the threshold for rejection using the Beta distribution (of course the explicit threshold can only be computed numerically). If the assumption of normality of the data is not satisfied, then this test is invalid. However, for large n as usual we can obtain an asymptotically level-α test.

Testing for independence in contingency tables: Here the measurements X and Y take values in $\{x_1, \ldots, x_k\}$ and $\{y_1, \ldots, y_\ell\}$, respectively. These x_i, y_j are categories, not numerical values (such as "smoking" and "non-smoking"). Let the total number of samples be n and let $N_{i,j}$ be the number of samples with values (x_i, y_j). Let $N_{i\cdot} = \sum_j N_{i,j}$ and let $N_{\cdot j} = \sum_i N_{i,j}$.

We want to test

$$H_0 : \ X \text{ and } Y \text{ are independent}$$

$$H_1 : \ X \text{ and } Y \text{ are not independent.}$$

Let $\mu(i,j) = \mathbf{P}\{X = x_i, Y = y_j\}$ be the joint pmf of (X,Y) and let $p(i)$, $q(j)$ be the marginal pmfs of X and Y respectively. From the sample, our estimates for these probabilities would be $\hat{\mu}(i,j) = N_{i,j}/n$ and $\hat{p}(i) = N_{i\cdot}/n$ and $\hat{q}(j) = N_{\cdot j}/n$ (which are consistent in the sense that $\sum_j \hat{\mu}(i,j) = \hat{p}(i)$ etc).

Under the null hypothesis we must have $\mu(i,j) = p(i)q(j)$. We test if these equalities hold (approximately) for the estimates. That is, define

$$T = \sum_{i=1}^{k} \sum_{j=1}^{\ell} \frac{(N_{i,j} - n\hat{p}(i)\hat{q}(j))^2}{n\hat{p}(i)\hat{q}(j)}.$$

Note that this is in the usual form of a χ^2 statistic (sum of (observed $-$ expected)^2/expected).

The number of terms is $k\ell$. We lose one d.f. as usual but in addition we estimate $(k-1)$ parameters $p(i)$ (the last one $p(k)$ can be got from the others) and $(\ell - 1)$ parameters $q(j)$. Consequently, the total degress of freedom is $k\ell - 1 - (k-1) - (\ell - 1) = (k-1)(\ell - 1)$.

Hence, we reject the null hypothesis if $T > \chi^2_{(k-1)(\ell-1)}(\alpha)$ to get (an approximately) level α test.

15. Regression and Linear regression

Let (X_i, Y_i) be i.i.d random variables. For example, we could pick people at random from a population and measure their height (X) and weight (Y). One question of interest is to predict the value of Y from the value of X. This may be useful if Y is difficult to measure directly. For instance, X could be the height of a person and Y could be the xxx

In other words, we assume that there is an underlying relationship $Y = f(X)$ for an unknown function f which we want to find. From a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ we try to guess the function f.

If we allow all possible functions, it is easy to find one that fits all the data points, i.e., there exists a function $f : \mathbb{R} \to \mathbb{R}$ (in fact we may take f to be a polynomials of degree n) such that $f(X_i) = Y_i$ for each $i \leq n$ (this is true only if we assume that all X_i are distinct which happens if X has a continuous distribution). This is not a good predictor, because the next data point (U,V) will fall way off the curve. We have found a function that "predicts" well all the data we have, but not for a future observation!

Instead, we fix a class of functions, for example the collection of all linear functions $y = mx + c$ where $m, c \in \mathbb{R}$ and within this class, find the best fitting function.

Remark 28. One may wonder if linearity is too restrictive. To some extent, but perhaps not as much as it sounds at first.

(1) Firstly, many relationships are linear in a reasonable range of the X variable (for example, resistance of a materiaal versus temperature).

(2) Secondly, we may sometimes transform the variables so that the relationship becomes linear. For example, if $Y = ae^{bX}$, then $\log(Y) = a' + b'X$ where $a' = \log(a)$ and $b' = \log(b)$ and hence in terms of the new variables X and $\log(Y)$, we have a linear relationship.

(3) Lastly, as a slight extension of linear regression, one can study *multiple linear regression*, where one has several independent variables $X^{(1)}, \ldots, X^{(p)}$ and try to fit a linear function $Y = \beta_1 X^{(1)} + \ldots + \beta_p X^{(p)}$. Once that is done, it increases the scope of curve fitting even more. For example, if we have two variable X, Y, then we can take $X^{(1)} = 1$, $X^{(2)} = X$, $X^{(3)} = X^2$. Then, linear regression of Y against $X^{(1)}, X^{(2)}, X^{(3)}$ is tantamount to fitting a quadratic polynomial curve for X, Y.

In short, multiple linear regression along with non-linear transformations of the individual variables, the class of functions f is greatly extended.

Finding the best linear fit: We need a criterion for deciding the "best". A basic one is the *method of least squares* which recommends finding α, β such that the error sum of squares $R^2 := \sum_{k=1}^n (Y_k - \alpha - \beta X_k)^2$ is minimized.

For fixed X_i, Y_i this is a simple problem in calculus. We get

$$\hat{\beta} = \frac{\sum_{k=1}^n (X_k - \bar{X}_n)(Y_k - \bar{Y}_n)}{\sum_{k=1}^n (X_k - \bar{X}_n)^2} = \frac{s_{X,Y}}{s_X^2}, \qquad \hat{\alpha} = \bar{Y}_n - \hat{\beta}\bar{X}_n$$

where $s_{X,Y}$ is the sample covariance of X, Y and s_X is the sample variance of X.

We leave the derivation of the least squares estimators by calculus to you. Instead we present another approach.

For a given choice of β, we know that the choice of α which minimizes R^2 is the sample mean of $Y_i - \beta X_i$ which is $\bar{Y} - \beta\bar{X}$. Thus, we only need to find $\hat{\beta}$ that minimizes

$$\sum_{k=1}^n \left((Y_k - \bar{Y}) - \beta(X_k - \bar{X})\right)^2$$

and then we simply set $\hat{\alpha} = \bar{Y} - \beta\bar{X}$. Let[4] $Z_k = \frac{Y_k - \bar{Y}}{X_k - \bar{X}}$ and $w_k = (X_k - \bar{X})^2 / s_X^2$. Then,

$$\sum_{k=1}^n \left((Y_k - \bar{Y}) - \beta(X_k - \bar{X})\right)^2 = s_X^2 \sum_{k=1}^n w_k \left(Z_k - \beta\right)^2.$$

Since w_k are non-negative numbers that add to 1, we can intepret it as a probability mass function and hence we see that the minimizing β is given by the expectation with respect to this mass

[4]We are dividing by $X_k - \bar{X}$. What if it is zero for some k? But note that in the expression $\sum \left((Y_k - \bar{Y}) - \beta(X_k - \bar{X})\right)^2$, all such terms do not involve β and hence can be safely left out of the summation. We leave the details for you to work out (the expressions at the end should involve all X_k, Y_k).

function. In other words,

$$\hat{\beta} = \sum_{k=1}^{n} w_k Z_k = \frac{s_{X,Y}}{s_X^2}.$$

Another way to write it is $\hat{\beta} = \frac{s_Y}{s_X} r_{X,Y}$ where $r_{X,Y}$ is the sample correlation coefficient.

A motivation for the least squares criterion: Suppose we make more detailed model assumptions as follows. Let X be a control variable (i.e., not random but we can tune it to any value, like temperature) and assume that $Y_i = \alpha + \beta X_i + \epsilon_i$ where ϵ_i are i.i.d. $N(0, \sigma^2)$ "errors". Then, the data is essential Y_i that are independent $N(\alpha + \beta X_i, \sigma^2)$ random variables. Now we can extimate α, β by the maximum likelihood method.

Example 29 (Hubble's 1929 experiment on the recession velocity of nebulae and their distance to earth)**.** Hubble collected the following data that I took from this website. Here X is the number of megaparsecs from the nebula to earth and Y is the observed recession velocity in 10^3km/s.

X	0.032	0.034	0.214	0.263	0.275	0.275	0.45	0.5	0.5	0.63	0.8	2
Y	0.17	0.29	-0.13	-0.07	-0.185	-0.22	0.2	0.29	0.27	0.2	0.3	1.09
X	0.9	0.9	0.9	0.9	1	1.1	1.1	1.4	1.7	2	2	2
Y	-0.03	0.65	0.15	0.5	0.92	0.45	0.5	0.5	0.96	0.5	0.85	0.8

We fit two straight lines to this data.

(1) Fit the line $Y = \alpha + \beta X$. The least squares estimators (as derived earlier) turn out to be $\hat{\alpha} = -0.04078$ and $\hat{\beta} = 0.45416$. If $Z_i = \alpha + \beta X_i$ are the predicted values of Y_is, then one can see that the *residual sum of squares* is $\sum_i (Y_i - Z_i)^2 = 1.1934$.

(2) Fit the line $Y = bX$. In this case we get $\hat{b}$ by minimizing $\sum_i (Y_i - bX_i)^2$. This is slightly different from before, but the same methods (calculus or the alternate argument we gave) work to give

$$\hat{b} = \frac{\sum_{i=1}^n Y_i X_i}{\sum_{i=1}^n X_i^2} = 0.42394.$$

The residual sum of squares $\sum_{i=1}^n (Y_i - bX_i)^2$ turns out to be 1.2064.

The residual sum of squares is smaller in the first, thus one may naively think that it is a better fit. However, note that the reduction is due to an extra parameter. Purely statistically, introducing extra parametrs will always reduce the residual sum of squares for obvious reasons. But the question is whether the extra parameter is worth the reduction. More precisely, if we fit the data too closely, then the next data point to be discovered (which may be nebula that is 10 megaparsecs away) may fall way off the curve.

More importantly, in this example, physics tells us that the line must pass through zero (that is, there is no recession velocity when two objects are very close). Therefore it is the second line

that we consider, not the first. This gives the Hubble constant to be 423 km./s./megaparsec (the currently accepted values appear to be about 70, with data going up to distances of hundreds of megaparsecs...see this data!).

Example 30. I have taken this example from the wonderful compilation of data sets by A. P. Gore, S. A. Paranjpe, M. B. Kulkarni. In this example, Y denotes the number of frogs of age X (in some delimited population).

X	1	2	3	4	5	6	7	8
Y	9093	35	30	28	12	8	5	2

A prediction about life-times says that the survival probability $P(t)$ (which is the chance that an individual survives up to age t or more) decays as $P(t) = Ae^{-bt}$ for some constants A and b. We would like to check this agains the given data.

What we need are individuals that survive beyond age t. Taking Z to be the cumulative sums of Y, this gives us

X	1	2	3	4	5	6	7	8
Z	9213	120	85	55	27	15	7	2
$P = Z/n$	1.0000	0.0130	0.0092	0.0060	0.0029	0.0016	0.0008	0.0002
$W = \log P$	0	-4.3409	-4.6857	-5.1210	-5.8325	-6.4203	-7.1825	-8.4352

We compute that $\bar{X} = 4.5$, $\bar{W} = -5.25$, $\text{std}(X) = 2.45$, $\text{std}(W) = 2.52$ and $\text{corr}(X, W) = 0.92$. Hence, in the linear regression $W = a + bX$, we see that $\hat{b} = 0.94$ and $\hat{a} = -9.49$. The residual sum of squares is 7.0.

How good is the fit? For the same data $(X_1, Y_1), \ldots, (X_n, Y_n)$, suppose we have two candidates (a) $Y = f(X)$ and (b) $Y = g(X)$. How to decide which is better? Or how to say if a fit is good at all?

By the least-squares criterion, the answer is the one with smaller residual sum of squares $SS := \sum_{k=1}^{n}(Y_k - f(X_k))^2$. Usually one presents a closely related quantity $R^2 = 1 - \frac{SS}{SS_0}$ (where $\text{SS}_0 = \sum_{k=1}^{n}(Y_k - \bar{Y})^2 = (n-1)s_Y^2$). Since SS_0 is (a multiple of) the total variance in Y, R^2 measures how much of it is "explained" by a particular fit. Note that $0 \leq R^2 \leq 1$. And higher (i.e., closer to 1) the R^2 is, the better the fit.

Thus, the first naive answer to the above question is to compute R^2 in the two situations (fitting by f and fitting by g) and see which is higher. But a more nuanced approach is preferable. Consider the same data and three situations.

(1) Fit a constant function. This means, choose α to minimize $\sum_{k=1}^{n}(Y_k - \alpha)^2$. The solution is $\hat{\alpha} = \bar{Y}$ and the residual sum of squares is SS_0 itself. Then, $R_0^2 = 0$.

(2) Fit a linear function. Then α, β are chosen as discussed earlier and the residual sum of squares is $\text{SS}_1 = \sum_{k=1}^{n}(Y_k - \hat{\alpha} - \hat{\beta}X_k)^2$. Then, $R_1^2 = 1 - \frac{SS_1}{SS_0}$.

(3) Fit a quadratic function. The the residual sum of squares is $\text{SS}_2 = \sum_{k=1}^{n}(Y_k - \hat{\alpha} - \hat{\beta}X_k - \hat{\gamma}X_k^2)^2$ where $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$ are chosen so as to minimize $\sum_{k=1}^{n}(Y_k - \alpha - \beta X_k - \gamma X_k^2)^2$. Then $R_2^2 = 1 - \frac{SS_2}{SS_0}$.

Obviously we will have $R_2^2 \geq R_1^2 \geq R_0^2$ (since linear functions include constants and quadratic functions include linear ones). Does that mean that the third is better? If that were the conclusion, then we can continue to introduce more parameters as that will always reduce the residual sum of squares! But that comes at the cost of making the model more complicated (and having too many parameters means that it will fit the current data well, but not future data!). When to stop adding more parameters?

Qualitatively, a new parameter is desirable if it leads to a *significant increase* of the R^2. The question is, how big an increase is significant. For this, one introduces the notion of *adjusted* R^2, which is defined as follows:

If the model has p parameters, then define $\bar{SS} = SS/(n-1-p)$. In particular, $\bar{SS}_0 = \frac{SS_0}{n-1} = s_Y^2$. Then define the adjusted R^2 as $\bar{R}^2 = 1 - \frac{\bar{SS}}{SS_0}$.

In particular, $\bar{R}_0^2 = R_0^2$ as before. But $R_1^2 = 1 - \frac{SS_1/(n-2)}{SS_0/(n-1)}$. Note that $\bar{R}^2$ does not necessarily increase upon adding an extra parameter. If we want a polynomial fit, then a rule of thumb is to keep adding more powers as long as $\bar{R}^2$ continues to increase and stop the moment it decreases.

Example 31. To illustrate the point let us look at a simulated data set. I generated 25 i.i.d $N(0,1)$ variables X_i and then generated 25 i.i.d. $N(0,1/4)$ variables ϵ_i. And set $Y_i = 2X_i + \epsilon_i$. The data set obtained was as follows.

X	-0.87	0.07	-1.22	-1.12	-0.01	1.53	-0.77	0.37	-0.23	1.11	-1.09	0.03	0.55
Y	-2.43	-0.56	-2.19	-2.32	-0.12	3.77	-1.4	0.84	0.34	1.83	-1.83	0.48	0.98
X	1.1	1.54	0.08	-1.5	-0.75	-1.07	2.35	-0.62	0.74	-0.2	0.88	-0.77	
Y	2.3	2.5	-0.41	-2.94	-1.13	-0.84	4.36	-1.14	1.45	-1.36	1.55	-2.43	

To this data set we fit two models (A) $Y = \beta X$ and (B) $Y = a + bX$. The results are as follows.

$$\text{SS}_0 = 96.20, \;\; R_0^2 = 0$$

$$\text{SS}_1 = 6.8651, \;\; R_1^2 = 0.9286, \;\; \bar{R}_1^2 = 0.9255$$

$$\text{SS}_2 = 6.8212, \;\; R_2^2 = 0.9291, \;\; \bar{R}_2^2 = 0.9227.$$

Note that the adjusted R^2 decreases (slightly) for the the second model. Thus, if we go by that, then the model with one parameter is chosen (correctly, as we generated from that model!). You

can try various simulations yourself. Also note the high value of R_1^2 (and R_2^2) which indicates that it is not a bad fit at all.