MATH 2B FINAL REVIEW 3/8/2012

ALDEN WALKER

1. Information

The final will be available on Thursday, March 8. It is due Wednesday, March 14, at noon. The time limit is 4 hours. You may use any sources from this class, including the books, course notes, TA notes, your notes, your homeworks, homework solutions, etc. You may use notes or other sources from anywhere else, like online course notes from another school. You may use a computer or calculator, but only in the ways in which you are allowed to use them on the homework. That is, for numerical CDF calculations, matrix algebra, etc. You can't use built-in regression programs or symbolic solvers, etc.

2. Outline

Here's what you should know about

- (Maximum likelihood) estimators
  - Bias
  - Mean squared error
- Hypothesis testing
  - Picking hypotheses
  - Finding statistics (Likelihood ratio - both from book and alternate)
  - Making a rejection rule
  - Find the constants in a rejection rule
  - Do the test
  - p-values
- Confidence intervals (and relationship to hypothesis testing)
  - For the mean of a normal using the t-distribution
  - The method of quantiles
- t-tests
  - Single
  - 2-population test
  - Paired t-test
- Regression
  - Picking a model
  - Getting least square estimates
  - Hypothesis testing and confidence intervals for:
    - β values
    - What we would get if we made another observation
- χ^2 tests
  - Goodness-of-fit
  - Independence
  - p-values for these

3. (Maximum likelihood) estimators

We are generally concerned with maximum likelihood estimators, but it's important to know the language about general estimators.

Date: 3/8/2012.

3.1. **Idea.** The idea is that you have some parameter (μ, say) on which the distribution of some random variables depend. You are given these random variables, and you want to estimate μ. The right way to do this is to find the value of μ which maximizes the probability of any observed data. The distribution of the X_i, thought of as a function of μ, is called the likelihood function.

Making it more concrete will be helpful, I think:

3.2. **Example (Discrete).** The discrete case is the most clear. Suppose that you know that you have $X_1 \dots X_n$ iid Poisson variables, and you want to find the mean of the Poisson distribution. Well, since they are independent, their joint density is a product of densities, i.e. $L(\lambda) = \prod_i \frac{e^{-\lambda}\lambda^{X_i}}{X_i!}$. We just now have to maximize this as a function of λ. It is usually easier to maximize the log of it when you are dealing with a product:

$$\log(L(\lambda)) = \sum_i (-\lambda) + X_i \log(\lambda) - \log(X_i!)$$

Taking the first derivative yields

$$\frac{d}{d\lambda} \log(L(\lambda)) = \sum_i -1 + X_i \frac{1}{\lambda}$$

We want this to be zero, i.e. $n = \frac{1}{\lambda}\sum_i X_i$, or rather, $\lambda = \frac{1}{n}\sum_i X_i$. This is not a huge surprise, but it is nice reassurance.

3.3. **Continuous Distributions.** These are the same, except the function that you optimize is the joint density functions. You solve them in the same way: (1) find the density function (perhaps it will be a joint density function if you have, as we did above, many trials (the X_i)) (2) think of it as a function of the parameter and (3) maximize it with respect to the parameter.

3.3.1. *Example.* I think you may have done this in class, but it never hurts to see it twice (the straightforward example I thought of (the exponential distribution) turned out to be on your homework).

Let's try to find the mle for the variance from n iid normal variables. Suppose that we know the mean μ. In practice, if you wanted to find μ and σ^2, you would first find the mle for μ because that won't depend on σ. Then you could use that to find σ. Anyway we know the mean is μ. Then the likelihood function is $L(\sigma^2) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp(\frac{-1}{2\sigma^2}(X_i - \mu)^2)$. We would then take logs to get $\log L(\sigma^2) = \sum_i -(1/2)\log(2\pi\sigma^2) + \frac{-1}{2\sigma^2}(X_i - \mu)^2$. The derivative of this is $\sum_i \frac{-1}{2\sigma^2} + \frac{1}{2\sigma^4}(X_i - \mu)^2$, so we should set $\sigma^2 = (1/n)\sum_i (X_i - \mu)^2$, as we expected.

3.4. **Definitions.** Of course, you want to be able to measure how good a job your estimator is doing. For that you would want the **mean squared error**. If $T(X)$ is your estimator for $g(\Theta)$, where Θ is the unknown parameter, then this is defined to be $E[(T(X) - g(\Theta))^2]$. It makes sense to take the expected value of this because it is a function of the random variable X (note that X may actually be a vector $(X_i)_i$ if you have multiple observations).

Another quantity that has a name is $E[T(X) - g(\Theta)]$. This is called the **bias**. If this is zero no matter what Θ is, we call the estimator **unbaised**.

Important mildy counterintuitive fact: it is not always true that the "best" estimator (i.e. the smallest MSE) is unbiased!

3.5. **Example 1.** It is always true that the sample mean $\bar{X_i}$ is an unbiased estimator of the true mean. This is because

$$\begin{aligned} E[(1/n)\sum_i X_i - \mu] &= (1/n)\sum_i E[X_i] - \mu \\ &= \mu - \mu = 0 \end{aligned}$$

However, the MSE of this estimator can change based on the distribution. For example, in the Poisson problem above,

$$\begin{aligned} E[((1/n)\sum_i X_i - \mu)^2] &= E[\frac{1}{n^2}\left(\sum_i X_i\right)^2 - 2(1/n)\mu\sum_i E[X_i] + \mu^2 \\ &= E[\left((1/n)\sum_i X_i\right)^2] - \mu^2 \\ &= E[Y^2] - E[Y]^2 \\ &= \mathrm{Var}(Y) \\ &= \frac{\lambda}{n} \end{aligned}$$

Where $Y = (1/n)\sum_i X_i$. Note that the variance of a sum (of ind vars) is the sum of the variances, and multiplying by a constant multiplies the variance by the square, so we are left with a $1/n$. This is good, since clearly the MSE of the sample mean should decrease with n. Also note that the MSE changes as the true value of λ changes.

3.6. **Example 2.** This is your homework problem, but it's a good example. Observations $X_1, \ldots, X_n$ are iid (independent and identically distributed) with exponential density

$$f_\theta(x) = \begin{cases} \frac{1}{\theta}\exp\left(\frac{-x}{\theta}\right) & \text{if } x > 0 \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is an unknown parameter.

Show that $\hat{\theta} = \overline{X}_n$ is the maximum likelihood estimator of θ and verify that this estimator is unbiased.

Since the X_i's are independent, the joint density function is

$$f(x_1, \ldots, x_n, \theta) = \prod_{i=1}^n \left(\frac{1}{\theta} + \exp\left(\frac{-x_i}{\theta}\right)\right)$$

when $x_i > 0$ for all i. Thus

$$\begin{aligned} \log f(x_1, \ldots, x_n, \theta) &= \sum_{i=1}^n \log\left(\frac{1}{\theta}\exp\left(\frac{-x_i}{\theta}\right)\right) \\ &= n\log\left(\frac{1}{\theta}\right) - \sum_{i=1}^n \frac{x_i}{\theta}. \end{aligned}$$

Differentiating this with respect to θ, we obtain

$$\frac{-n}{\theta} + \sum_{i=1}^n \frac{x_i}{\theta^2}.$$

Setting this equal to 0 and multiplying through by θ^2, we have

$$-n\theta + \sum_{i=1}^n x_i = 0,$$

and hence

$$\theta = \frac{\sum_{i=1}^n x_i}{n}.$$

Thus we have shown that $\hat{\theta} = \overline{X}_n$. The expected value of this estimator is

$$\begin{aligned} E(\overline{X}_n) &= E\left(\frac{\sum_{i=1}^n X_i}{n}\right) \\ &= \frac{E\sum_{i=1}^n (X_i)}{n} \\ &= E(X) \\ &= \int_0^\infty \frac{x}{\theta}\exp\left(\frac{-x}{\theta}\right) dx \\ &= \left(-x\exp\left(\frac{-x}{\theta}\right)\right)_0^\infty + \int_0^\infty \exp\left(\frac{-x}{\theta}\right) dx \\ &= \left(-\theta\exp\left(\frac{-x}{\theta}\right)\right)_0^\infty \\ &= \theta. \end{aligned}$$

Thus $\overline{X}_n$ is an unbiased estimator for θ.

Calculate the Mean Square Error $E_\theta(c\hat{\theta}-\theta)^2$ for $c > 0$ of a modified estimator $c\hat{\theta}$ and determine the value of c that minimizes it. (Hint: OK to use the facts that each observation has mean θ and variance θ^2.)

The mean square error of $c\hat{\theta}$ is given by

$$\begin{aligned} E\big((c\hat{\theta}-\theta)^2\big) &= E\big(c^2\hat{\theta}^2 - 2c\hat{\theta}\theta + \theta^2\big) \\ &= c^2E(\hat{\theta}^2) - 2c\theta E(\hat{\theta}) + \theta^2. \end{aligned}$$

Therefore to compute this we want to find $E(\hat{\theta})$ and $E(\hat{\theta}^2)$. We know from (a) that $E(\hat{\theta}) = \theta$, so therefore we only need to calculate $E(\hat{\theta}^2)$. To do this, first observe that, since we know that $Var(X) = E(X^2) - (E(X))^2$, the hint implies that $E(X^2) = 2\theta^2$. Therefore we compute that

$$\begin{aligned} E\big(\hat{\theta}^2\big) &= E\Big(\overline{X}_n^2\Big) \\ &= E\left(\left(\frac{\sum_{i=1}^n X_i}{n}\right)^2\right) \\ &= E\left(\frac{\sum_{i=1}^n X_i^2 + \sum_{i\neq j} X_iX_j}{n^2}\right) \\ &= \frac{\sum_{i=1}^n E(X_i^2) + \sum_{i\neq j} E(X_i)E(X_j)}{n^2} \\ &= \frac{\sum_{i=1}^n 2\theta^2 + \sum_{i\neq j}\theta^2}{n^2} \\ &= \frac{2n\theta^2 + n(n-1)\theta^2}{n^2} \\ &= \frac{(n^2+n)\theta^2}{n^2} \\ &= \frac{n+1}{n}\theta^2. \end{aligned}$$

Thus the mean square error of $c\hat{\theta}$ is

$$\begin{aligned} E\big((c\hat{\theta}-\theta)^2\big) &= c^2E(\hat{\theta}^2) - 2c\theta E(\hat{\theta}) + \theta^2 \\ &= \frac{(n+1)c^2}{n}\theta^2 - 2c\theta^2 + \theta^2 \\ &= \frac{(n+1)c^2 - 2cn + n}{n}\theta^2. \end{aligned}$$

To minimize this, it suffices to minimize $(n+1)c^2 - 2cn + n$. Since this is a quadratic in c with positive leading coefficient, it should have a global minimum at a point where the derivative of this expression with

respect to c equals 0. Setting the derivative equal to 0, we obtain

$$2c(n+1) - 2n = 0,$$

and hence

$$c = \frac{n}{n+1}.$$

Thus the estimator of the form $c\hat{\theta}$ which minimizes mean square error is $\frac{n}{n+1}\hat{\theta}$.

4. Hypothesis testing

Just like with estimators, we are going to try to figure something out about the parameters of a distribution just from observing the distribution. In this setup, we pick two hypotheses which completely fill the possible parameter space. In other words, we know the mean is either 0 or 1, so we pick the hypotheses "it's 1" and "it's 0". Or perhaps we don't know the mean at all, but we're interested in 0, so we pick "it's 0" and "it's not 0".

However, the situation isn't symmetric: there is a special ("null") hypothesis on which the test is focused, so we will either accept or reject the null hypothesis.

4.1. Some Definitions.

You pick a null hypothesis H_0:

- The probability of rejecting H_0 when it is true is the **significance level**, usually denoted α.
- If the probability of accepting H_0 when it is false is β, then the **power** of the test is $1 - \beta$ (the probability of rejecting it when it is, in fact, false). It is common to ask about the power of the test "at" a value of the parameter you are estimating. This means that assuming the true value of the parameter is given (and it's not in the null hypothesis), calculate the probability of rejecting the null (and take one minus that).
- The test **statistic** is the number that you compute from your data (like the sample mean, for example), and determining whether or not that number lies in a certain range is your "test."
- To "determine critical values" for a test with a given test statistic T and significance α means to find the values of T which should indicate acceptance or rejection of the null hypothesis in such a way that the probability of rejecting the null hypothesis when it is true is α.
- To choose which hypothesis gets to be the null hypothesis, you pick the one that you would want to favor when you are unsure. For example, typically you set α quite small. This means that near the null hypothesis you will default to the null hypothesis. That is, it is the one that you assume unless it is proven wrong.

  It is often the case that you want to choose the "bad" thing as the null hypothesis because the consequences of incorrectly assuming something is good are worse. For example, if you are testing a new drug to see if it is safe, the null hypothesis is: "it's not safe." That way, you can control (and make small) the probability of incorrectly rejecting the null (saying it's safe incorrectly). This is clearly better because you want to be really sure that your new drug incur any lawsuits and such.

4.2. Recipe.

(1) Get a parameter you are supposed to draw conclusions about
(2) Choose hypotheses and a significance level
(3) Choose a test statistic (likelihood ratio?)
(4) Choose a rejection rule
(5) Determine the constants in the rejection rule such that the significance level is what it's supposed to be.

4.3. Example ((α, β) Specification).

You always want to control the significance α, but sometimes you want to set a value of the parameter in the alternative hypothesis space and say that if the true value is close to this chosen value, you want to make sure to reject the null. Let's see an example:

Let's say we have the distribution

$$f_\Theta(x) = \begin{cases} \Theta x^{\Theta-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

And we want to find Θ, or at least draw some conclusions about it. Our null hypothesis will be $\Theta < \Theta_0$, so H_1 is "$\Theta \geq \Theta_0$". Suppose that we want a significance level of $\alpha = 0.05$, but we also want the probability of

accepting the null (incorrectly) when $\Theta = \Theta_1 > \Theta_0$ to be small (say $\beta = 0.05$). This is our α, β specification. How many samples do we need to take, and what should our critical point be? Also, what should our statistic be?

In this case, the likelihood ratio is (note we've picked $\Theta_0 = \mu'$ and $\Theta_1 = \mu''$ to match it up to the definition above):

$$\prod_i \frac{\Theta_0 X_i^{\Theta_0 - 1}}{X_i^{\Theta_1 - 1}\Theta_1} = \left(\frac{\Theta_0}{\Theta_1}\right)^n \left(\prod_i X_i\right)^{\Theta_0 - \Theta_1}$$

This uses the alternate formulation above.

Note this depends bijectively with simply $\prod_i X_i$, so that is our test statistic. However, since $\Theta_1 > \Theta_0$, it's an orientation-reversing bijection. That is, we will reject the null if our statistic is *larger* than a constant C. Unfortunately, the distribution of $\prod_i X_i$ isn't too easy. However, now let's take the log and define our test function $T(X_i) = \sum_i \log(X_i)$. Of course, we haven't made the distribution easier, but now we can use the normal approximation (assuming n is fairly large). We need to know the mean and variance of $\log(X_i)$. The mean is $\int_0^1 \log(x)\Theta x^{\Theta-1} dx$. You can do this fairly easily with integration by parts and get $E(\log(X_i)) = -1/\Theta$. We can also compute $E(\log(X_i)^2) = \int_0^1 \log(x)^2 \Theta x^{\Theta-1} dx$, which comes out to $2/\Theta^2$. Therefore,

$$Var(\log(X_i)) = E(\log(X_i)^2) - E(\log(X_i))^2 = \frac{2}{\Theta^2} - \frac{1}{\Theta^2} = \frac{1}{\Theta^2}$$

Therefore, we can approximate the distribution of $\sum_i \log(X_i)$ by a normal $(n\left(\frac{-1}{\Theta}\right), n\frac{1}{\Theta^2})$ distribution. This holds for any Θ, so we now have two equations: for α, we have $P(\text{reject}|\text{null}) = P(\sum_i \log(X_i) > C|\text{null}) < 0.05$, and from β we get $P(\sum_i \log(X_i) < C|\ \Theta = \Theta_1) < 0.05$. This gives two equations in the unknowns n and C, which you can solve for, given any two particular Θ values.

To make this concrete, suppose $\Theta_0 = 1/4$ and $\Theta_1 = 3/4$. Then the null distribution is (approximately) $N(-4n, 16n)$, and the distribution for Θ_1 is $N(-(4/3)n, (16/9)n)$. Therefore

$$P(\sum_i \log(X_i) > C|\text{null}) \approx 1 - \Phi\left(\frac{C + 4n}{4\sqrt{n}}\right)$$

and

$$P(\sum_i \log(X_i) < C|\ \Theta = \Theta_1) \approx \Phi\left(\frac{C + (4/3)n}{(4/3)\sqrt{n}}\right)$$

Note that you will have to use the inverse of Φ here. Now you can either just go and experimentally try it, but a better way is to use Φ^{-1} to get two equalities for C, and then solve for n. That is, from the first and second equations, respectively, we get:

$$C = (4\sqrt{n})\Phi^{-1}(1 - 0.05) - 4n$$

and

$$C = (4/3)\sqrt{n}\Phi^{-1}(0.05) - (4/3)n$$

Setting these equal we get $(4 + (4/3))\Phi^{-1}(0.95) = (4 - (4/3))\sqrt{n}$, which gives us $n = (2\Phi^{-1}(0.95))^2 = 10.82$, which means that we need 11 trials to be able to satisfy this (C turns out to be about -21.6).

5. t-Tests

Again, the notes on the course website are highly recommended.

These tests arise when you have a data sample, and you assume it's normal, but you know neither the mean nor the variance. Suppose you want to estimate the mean. Well, if you did know the standard deviation, then you would use the test statistic $\frac{\bar{X}_i - \mu_0}{\sigma/\sqrt{n}}$, which would be normal, and everything would be good. However, if you don't know the standard deviation, then you use the obvious replacement—the sample standard deviation $s = \sqrt{\frac{\sum_i (X_i - \bar{X})^2}{n-1}}$. Because of the variability of the sample standard deviation, though, that statistic is no longer normal. It has what is called a t-distribution with $n - 1$ degrees of freedom.

5.1. **Basic.** You just use the statistic

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

Which, under the null hypothesis that $\mu = \mu_0$, will have a t distribution with $n-1$ degrees of freedom.

5.2. **2 Population t-Tests.** This is a similar situation, but here we have two populations which are normal with the same standard deviation and different means. Then to test hypotheses involving the difference of the means, we can use the statistic

$$T = \frac{\bar{X}_i - \bar{Y}_i - (\mu_X - \mu_Y)}{s\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

Where

$$s = \sqrt{\frac{\sum_i (X_i - \bar{X}_i)^2 + \sum_i (\bar{Y}_i - Y_i)^2}{m+n-2}}$$

Where there are n X_i's and m Y_i's. The statistic T has a t-distribution with $n+m-2$ degrees of freedom, so you can use a similar method to the above to set up tests. The setup and stuff is virtually identical once you have the distribution.

5.3. **Paired t-Tests.** Here you have a set of data pairs, and you want to evaluate the average differences. This does not require the assumption that the values in the pairs are normal, but only that the differences are normal. Here you will use the statistic

$$T = \frac{(\bar{X}_i - \bar{Y}_i)}{\frac{1}{\sqrt{n}}\sqrt{\frac{\sum_i \left((X_i - Y_i) - (\bar{X}_i - \bar{Y}_i)\right)^2}{n-1}}}$$

Which has a t distribution with $n-1$ degrees of freedom. Note this is basically just a one population t-test, applied to the differences.

5.4. **Example.** Suppose that I have 10 people, and I do the following experiment. During the first test, participants play a video game for a few minutes and are then tested for reaction time. During the second test, they don't play a video game but are still tested for reaction time. I get the following reaction times that I just made up (each person is a row).

game	no game
0.1	0.3
0.2	0.3
0.3	0.34
0.11	0.28
0.33	0.4
0.16	0.14
0.10	0.22
0.13	0.12
0.13	0.20
0.14	0.13

Does playing a video game improve your reaction time? Let's see.

5.4.1. *2-population test.* Let's do a two-population t-test to compare the means for the "game" population vs the "no game" population at the 0.05 significance level. Our hypotheses are:

$$H_0 : \mu_X = \mu_Y \text{ vs. } H_1 : \mu_X \neq \mu_Y$$

So we set $\mu_X = \mu_Y$ in the 2-population t statistic above, and we have

$$S_p^2 = \frac{\sum_{i=1}^{10}(X_i - \bar{X})^2 + \sum_{i=1}^{10}(Y_i - \bar{Y})^2}{10+10-2} = 0.00798944$$

and

$$T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{10} + \frac{1}{10}}} = \frac{0.17 - 0.243}{\sqrt{0.00798944}\sqrt{1/5}} = -1.82626$$

Under the null, T has a t distribution with $10 + 10 - 2 = 18$ degrees of freedom, and our rejection rule is "reject if $|T| \geq C$". We choose C such that $P(|T| \geq C \,|\, \mu_X = \mu_Y) = 0.01$), so $C = |t_{18,0.025}| = 2.10$. Notice that our statistic is -1.82626, so we do *not* reject the null, and we conclude that playing a video game does not lower your reaction time

5.4.2. *Paired test.* Now let's do a paired t-test at the same significance level of 0.05. We will assume the difference between the times for each person are normal, and we want to know the mean (is the mean = 0?). Our hypotheses are $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$, where μ here is the true mean of the *difference* between each pair. Under the null hypothesis, our statistic

$$T = \frac{(\bar{X}_i - \bar{Y}_i)}{\frac{1}{\sqrt{n}}\sqrt{\frac{\sum_i \left((X_i - Y_i) - (\bar{X}_i - \bar{Y}_i)\right)^2}{n-1}}} = -3.03669$$

Has a t distribution with 9 degrees of freedom. We reject if $|T| \geq C$, and we choose C so that $P(|T| \geq C \,|\, \mu = 0) = 0.01$. This means that $C = |t_{9,0.025}| = 2.26$.

In fact, our statistic is not in this interval, so we reject the null; playing a video game does alter your reaction time. Clearly, it reduces the time, and we would also reject the null hypothesis that the mean of the "game - no game" distribution is nonnegative.

Thus, with the two-population t-test, we failed to conclude that playing the game is helpful, while the paired t-test did conclude there was a difference. Of course, if we were less strict, we might have gotten the same conclusion.

There is also a difference in the assumptions you must make: the two-population t-test assumes the populations are normal, while the paired t-test assumes the differences are normal.

5.4.3. *Further example.* Suppose I wanted to test the hypothesis that playing the video game multiplied your reaction time by 0.75. Then I could do a paired t-test on $X - 0.75Y$. Here my statistic is:

$$T = \frac{(\bar{X}_i - 0.75\bar{Y}_i)}{\frac{1}{\sqrt{n}}\sqrt{\frac{\sum_i \left((X_i - 0.75Y_i) - (\bar{X}_i - 0.75\bar{Y}_i)\right)^2}{n-1}}} = -0.590514$$

This statistic is quite small in aboslute value compared to the constants, so we cannot reject the null that playing the game reduces your reaction time to 75% of its previous value.

Why was I allowed to do this weird 0.75 **thing?** The reason is because I simply changed my assumptions! In the paired test, I assumed that the differences were normal. Now I'm assuming that the difference between X and $0.7Y$ is normal!

It's always important to keep in mind what your assumptions are.

6. p-VALUES

In the above example, we rejected using a paired t test but failed to reject using a 2-population t test. This indicates that the paired t test is more powerful, and if you think of some examples, it's clear why it might be more discerning. However, note that if we set, for example, $\alpha = 0.01$, we would not have rejected for either test.

Somehow in real life you don't really want to know whether you accept or reject at a certain significance level — you want to know whether you accept or reject for any given significance level. This is where the p-value comes it.

The p-value of a statistic is the propability of observing something as or more extreme under the null hypothesis. For example, with our 2-population t test above, it's the probability of observing something with $|T| \geq 1.826$ under the null hypothesis that $\mu_X = \mu_Y$. Since T has a t distribution, we can just calculate this: it's $p = 0.0844$.

Similarly, for the paired t test, we want the probability under the null that $|T| \geq 3.037$, which is $p = 0.014$. Note that you will reject the null for any significance level α such that $\alpha \geq p$.

7. Confidence Intervals

The idea is that, given some statistic (observations), you want to give an interval such that with probability α, the true value of a parameter lies within the interval.

The way that you do all of these follows the same idea. You somehow write down what you want, or what you know, in terms of probabilities, and then fiddle with it until you get what you want.

7.1. **The mean of a normal.** Let's give a 90% confidence interval for the mean of a normal distribution, given 20 samples from that distribution with sample mean 10 and variance 5.

Ok: start with what we know. The statistic $\frac{\bar{X}_i-\mu}{s/\sqrt{n}}$, where μ is the true mean and s is the sample standard deviation, has a t distribution with $n-1$ degrees of freedom, so we know that

$$P\left(t_{n-1,0.05} \leq \frac{\bar{X}_i - \mu}{s/\sqrt{n}} \leq t_{n-1,0.95}\right) = 0.9$$

Rearranging, we know that

$$P\left(\bar{X}_i - |t_{n-1,0.05}|\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X}_i + |t_{n-1,0.95}|\frac{s}{\sqrt{n}}\right) = 0.9$$

And thus with probability 0.9, the true mean lies within the interval with endpoints $\bar{X}_i \pm |t_{n-1,0.05}|\frac{s}{\sqrt{n}}$. Note that the t distribution is symmetric.

In our example, then, our 90% interval is $10 \pm |t_{19,0.05}|\frac{\sqrt{5}}{\sqrt{20}} = 10 \pm 0.865$

7.2. **The method of quantiles.** Suppose you have a generic distribution, like Poisson(λ), and you'd like a confidence interval for λ. I will continue this discussion in the case of Poisson, but it's the same for any distribution. How do you do it? Suppose you have a statistic X obtained from the data. Then a $(1-\alpha)$ confidence interval is given by $[\underline{\lambda}, \overline{\lambda}]$, where

$$P(Y > X \,|\, \lambda = \underline{\lambda}) = \alpha/2 \quad \text{and} \quad P(Y < X \,|\, \lambda = \overline{\lambda}) = \alpha/2$$

Where Y is a random variable *drawn from the same distribution as* X. Here you think of X as fixed, and Y is the random variable. You want to figure out what $\overline{\lambda}$ and $\underline{\lambda}$ should be to satisfy the constraints.

Let's finish this example. Suppose we have sampled our Poisson distribution with 50 samples, and we find that the mean is 4.52. What's a 95% confidence interval for λ? We set our statistic X to be the mean 4.52.

We want to find $\underline{\lambda}$ and $\overline{\lambda}$ so that

$$P(Y > 4.52 \,|\, \lambda = \underline{\lambda}) = 0.025 \quad \text{and} \quad P(Y < 4.52 \,|\, \lambda = \overline{\lambda}) = 0.025$$

Where Y is a random variable which is the mean of 50 iid Poisson random variables with parameter λ.

This is somewhat difficult to solve exactly, so let's use the normal approximation. We know that the variance of a Poisson(λ) random variable is λ. Therefore, Y has mean λ and variance λ/n. Here $n = 50$, but I think it's clearer if I leave it symbolically in for now. By the central limit theorem, Y has approximately a Normal($\lambda, \lambda/n$) distribution. Therefore, we want

$$0.025 = P(Y > 4.52 \,|\, \lambda = \underline{\lambda}) = P\left[\frac{Y-\lambda}{\sqrt{\lambda/n}} > \frac{4.52-\lambda}{\sqrt{\lambda/n}} \,\middle|\, \lambda = \underline{\lambda}\right]$$

But the thing on the left inside the probability on the right is a standard normal variable! Thus, we want to find $\underline{\lambda}$ such that

$$1 - \Phi\left[\frac{4.52-\underline{\lambda}}{\sqrt{\underline{\lambda}/n}}\right] = 0.025$$

Similarly, we want

$$\Phi\left[\frac{4.52-\overline{\lambda}}{\sqrt{\overline{\lambda}/n}}\right] = 0.025$$

Simplifying,

$$\frac{4.52-\underline{\lambda}}{\sqrt{\underline{\lambda}/50}} = \Phi^{-1}(0.975) = 1.95996 \quad \text{and} \quad \frac{4.52-\overline{\lambda}}{\sqrt{\overline{\lambda}/n}} = \Phi^{-1}(0.025) = -1.95996$$

Squaring both sides gives us a single quadratic equation, whose two solutions will then be our $\underline{\lambda}$ and $\overline{\lambda}$. In this case, we get

$$\underline{\lambda} = 3.96787 \quad \text{and} \quad \overline{\lambda} = 5.14896$$

This gives our 95% confidence interval. We used the central limit theorem, so it's approximate, but as n gets large, it will be very accurate.

Just to satisfy my curiousity, I generated 50 Poisson(5) samples and computed a 95% confidence interval in this way. I did this $10,000$ times, and my interval succeeded (contained 5) 9,483 times, so this is probably working.

8. Regression!

Suppose we are fitting the linear model $y = \beta_1 x + \beta_2 + \epsilon$, where ϵ is normal with mean 0 and variance σ^2. In order to try to find β_1 and β_2, we find them so has to solve the following equation "as well as possible".

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

(Denote the matrix on the right by M).

Well, of course you can't do that perfectly because the system is very overdetermined. However, you can find β such that the right hand side is as close as possible to the $\mathbf{y}$ vector. Note that this closest vector will clearly (if you think about it) be the projection of $\mathbf{y}$ onto the column space of M. And, it turns out there's a nice formula for this projection!

There are two relevant formulas here; the formula for the *projection* $\hat{\mathbf{y}}$ is $\hat{\mathbf{y}} = M(M^TM)^{-1}M^T\mathbf{y}$. The formula for the *coefficients* of the projection in terms of the basis given by the columns of M (this is what you will want in this senario) is then $\hat{\beta} = (M^TM)^{-1}M^T\mathbf{y}$. Note that if you were doing this over $\mathbb{C}$ rather than $\mathbb{R}$, you would have adjoints instead of transposes (the matrix is $M(M^*M)^{-1}M^*$).

The difference between the vector $\mathbf{y}$ and its projection $\hat{\mathbf{y}}$ (i.e. the length) is the SSR (sum of the squares of the residuals). You could write it as $\|\mathbf{y} - \hat{\mathbf{y}}\|^2$.

8.1. Facts.

- $\hat{\mathbf{y}} = M(M^TM)^{-1}M^T\mathbf{y}$
- $\hat{\beta} = (M^TM)^{-1}M^T\mathbf{y}$.
- $SSR = \|\mathbf{y} - \hat{\mathbf{y}}\|^2$.
- Suppose that you have n observations, and r regression coefficients (the length of β). Then it turns out that
$$\frac{c\beta - c\hat{\beta}}{s\sqrt{c(M^TM)^{-1}c^T}} \sim t_{n-r}$$
where: "$\sim$" means "is distributed as," $s = \sqrt{\frac{SSR}{n-r}}$, and c is any vector of length r (like, if you're asking about β_0, you could use $c = (1, 0, \ldots, 0)$).
- And $E(c\hat{\beta}) = c\beta$, which means $c\hat{\beta} \sim N(c\beta, c(M^TM)^{-1}c^T\sigma^2)$.
- The SSR is also called the RSS, and we can use it to produce an unbiased estimator of the error variance σ^2, i.e. $\widehat{\sigma^2} = s^2 = \frac{\|\mathbf{y}-\hat{\mathbf{y}}\|^2}{n-r}$
- Another useful fact is that $Var(c\hat{\beta}) = c(M^TM)^{-1}c^T\sigma^2$, and we use this to make the definition of the *standard error* of an estimate $c\hat{\beta}$ as $s\sqrt{c(M^TM)^{-1}c^T}$. The standard error of one of the coefficients β_i is just the standard error of $c\hat{\beta}$ where c has a 1 in the ith spot and zeros elsewhere.
- You can also guess a new observation. Suppose you'd like to know $\mathbf{y}_{n+1} = c\beta$. The obvious guess is $c\hat{\beta}$, and the question is: how good a guess is this? Well, it turns out that
$$\frac{\mathbf{y}_{n+1} - c\hat{\beta}}{s\sqrt{1 + c(M^TM)^{-1}c^T}} \sim t_{n-r}$$
From which you can get confidence intervals or test hypotheses.

This essentially tells you all you need to know to give confidence intervals and deal with hypotheses involving β.

8.2. **Example.** Let's fit a line $\beta_1 x + \beta_2$ to the following data and give 95% confidence intervals for the β_i:

y	x
0.801033	0
4.23744	1
5.48164	2
3.20569	3
6.24067	4
7.45512	5
8.21995	6
10.3746	7
9.10986	8
11.861	9
12.25	10

The matrix M here is:

$$M = \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 2 & 1 \\ 3 & 1 \\ 4 & 1 \\ 5 & 1 \\ 6 & 1 \\ 7 & 1 \\ 8 & 1 \\ 9 & 1 \\ 10 & 1 \end{bmatrix}$$

Therefore we compute

$$\hat{\beta} = (M^T M)^{-1} M^T \mathbf{y} = \begin{bmatrix} 1.04492 \\ 1.97878 \end{bmatrix}$$

and

$$\hat{\mathbf{y}} = M\hat{\beta} = \begin{bmatrix} 1.97878 \\ 3.0237 \\ 4.06862 \\ 5.11353 \\ 6.15845 \\ 7.20337 \\ 8.24828 \\ 9.2932 \\ 10.3381 \\ 11.383 \\ 12.428 \end{bmatrix}$$

I did this in Mathematica (I will put the notebook on my website).

The $SSR = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 = 11.506$

Let's get confidence intervals: first, let $c = (1, 0)$. We know that $\frac{c\beta - c\hat{\beta}}{s\sqrt{c(M^T M)^{-1}c^T}} \sim t_{n-r}$, where r here is 2, and $s = \sqrt{SSR/(n-r)} = \sqrt{11.506/(11-2)} = 1.13$. Using Mathematica (or noting that $c(M^T M)^{-1}c^T$ is just the upper left element of $(M^T M)^{-1}$), we find that $\sqrt{c(M^T M)^{-1}c^T} = 0.0953$, and thus

$$\begin{aligned} \frac{c\beta - c\hat{\beta}}{1.13 \times 0.0953} \sim t_{n-r} \quad &\Rightarrow \quad P\left(t_{9,0.025} \leq \frac{\beta_1 - \hat{\beta}_1}{0.107806} \leq t_{9,0.975} \right) = 0.95 \\ &\Rightarrow \quad P\left(\hat{\beta}_1 + (0.107806) t_{9,0.025} \leq \beta_1 \leq \hat{\beta}_1 + (0.107806) t_{9,0.975} \right) = 0.95 \end{aligned}$$

And thus a 95% confidence interval for β_1 is $\hat{\beta}_1 \pm (0.107806)t_{9,0.975}$. Plugging in values, this is $1.04492 \pm (0.107806)t_{9,0.975} = [0.801042, 1.28879]$.

Following an identical calculation, we find a 95% confidence interval for β_2 of $[0.6219, 3.336]$. Thinking about this for a moment in the context of this being an example problem, you can guess that $\beta_1 = 1$, and β_2 is 1, 2, or 3, and in fact, checking my Mathematica notebook shows how I generated the data and that this is true.

We can also get the standard error of, for example, $\hat{\beta}_1 + \hat{\beta}_2$. We know

$$s^2_{\hat{\beta}_1+\hat{\beta}_2} = c(M^TM)^{-1}c^Ts^2 = 0.2363 \times 1.13 = 0.267$$

Where $c = (1, 1)$.

Let's calculate how good a fit this is, generally. We calculate the coefficient of multiple determination:

$$R^2 = \frac{s_y^2 - s_{\hat{\epsilon}}^2}{s_y^2} = \frac{13.161 - 1.15}{13.161} = 0.91258$$

So that's a pretty good fit.

Now suppose that we make another observation $\mathbf{y}_{n+1}$ at 11, so set $c = (11, 1)$. Now we know

$$\frac{\mathbf{y}_{n+1} - c\hat{\beta}}{s\sqrt{1 + c(M^TM)^{-1}c^T}} \sim t_{n-r}$$

So we compute $c\hat{\beta} = 13.47$ and $s\sqrt{1 + c(M^TM)^{-1}c^T} = 1.3457$, so we know

$$P\left(t_{9,0.025} \leq \frac{\mathbf{y}_{n+1} - 13.47}{1.3457} \leq t_{9,0.975}\right) = 0.95$$

And thus a 95% confidence interval for the new observation is:

$$13.47 \pm 1.3457t_{9,0.975} = [10.426, 16.514]$$

9. χ^2 tests

9.1. **Goodness-Of-Fit Tests.** You can use a χ^2 test to see if data is distributed according to a given distribution. The assumption here are that your data is in bins. Then you do the following.

- Let the null hypothesis be that the data is normal (or whatever distribution)
- Compute the expected number E_i for each bin, given that the distribution is normal (or whatever) with the sample mean and stddev (or the given mean and stddev, in the case of question 5)
- Find the $\sum_{i=0}^{n}(O_i - E_i)^2/E_i$, where O_i is the observed number in the bin i.
- That sum is the chi-square statistic X^2, and we assume/it is distributed as a chi-square distribution with $n - m - 1$ degrees of freedom, where m is the number of parameters being estimated (in the normal case, 2)
- Find the probability that a chi-square random variable with $n - m - 1$ degrees of freedom is larger than X^2. This is the p-value, and thus for any $\alpha > p$, we reject the null hypothesis that the data is normal.

9.1.1. *Example (goodness-of-fit).* Suppose you have a bunch of bacteria clumps, and you think that the number of clumps per square is Poisson. Well, let's use a χ^2 test. Here is the data:

Number per square	0	1	2	3	4	5	6	7	8	9	10	19
Frequency	56	104	80	62	42	27	9	9	5	3	2	1

The sample mean here is just the average $(0 \times 56 + 1 \times 104 + \cdots + 19 \times 1)/400 = 2.44$. Under this mean, we would expect the following counts (for example, we expect $400 \times e^{-\lambda} = 34.9$ in bin 0:

Number per square	0	1	2	3	4	5	6	≥ 7
Observed	56	104	80	62	42	27	9	20
Expected	34.9	85.1	103.8	84.4	51.5	25.1	10.2	5.0
$\frac{(O-E)^2}{E}$	12.8	4.2	5.5	5.9	1.8	0.14	0.14	45.0

Notice that we have binned the data so that each bin has at least 5 items.

The total statistic is $X^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 75.4$. We have 6 degrees of freedom because there are 8 bins and we are estimating one parameter (the mean of the Poisson), so $8 - 1 - 1 = 6$. Then we find that if $Y \sim \chi^2_6$, we have $P(Y > 75.4) \approx 3.1 \times 10^{-14}$, so pretty much there's no chance that the data is Poisson. This is interesting because looking at a plot of the data, you might think it was. You can see how it fails the test by looking that the bottom row of the table—notice that the main failure is that there were too many observed counts that were high and low.

We could do the same test, but we could simply check whether a given fixed distribution fits the data; that is, we don't estimate λ, we just say "let's test if this is Poisson with mean 2.5. In this case, we have:

Number per square	0	1	2	3	4	5	6	≥ 7
Observed	56	104	80	62	42	27	9	20
Expected	32.83	82.08	102.60	85.50	53.44	26.72	11.13	5.674
$\frac{(O-E)^2}{E}$	16.34	5.850	4.980	6.461	2.449	0.00292	0.408	36.16

In this case the sum is 72.65. We haven't estimated any parameters from the data, so we have $8 - 1 = 7$ degrees of freedom. This gives a p-level of `1-CDF[ChiSquareDistribution[7], 72.65]` $= 4.27 \times 10^{-13}$. So yeah that's pretty unlikely.

9.2. **Independence.** The test for independence is really the same, except it's slightly tricky to calculate the expected number. What you should do is to put your data into a table, like:

	child low bp	mid bp	high bp
father low bp	14	11	8
mid bp	11	11	9
high bp	6	10	12

Then to test for independence, you compute the statistic:

$$X^2 = \sum_i \frac{\left(n_{ij} - \frac{n_{i\cdot} n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot} n_{\cdot j}}{n}}$$

This is just the statistic above, where $n_{i\cdot} n_{\cdot j}/n$ is the expected number. Think about why that might be the right number. The notation $n_{i\cdot}$ means the sum of the entries in the ith row, and $n_{\cdot j}$ means the sum of the entries in the jth column. This statistic has $(I-1)(J-1)$ degrees of freedom, where I and J are the number of rows and columns, respectively. Plugging in the numbers in the matrix above, we get

$$X^2 = 3.81436$$

Since $I = J = 3$, we have 3 degrees of freedom. Our p-value is `1-CDF[ChiSquareDistribution[3],3.81436]` $= 0.28222$. This isn't that small, so we probably don't reject the hypothesis that the father's blood pressure is independent of the child.