Lecture 8: Inference on Parameters

36-401, Fall 2015, Section B

24 September 2015

Contents

Having gone over the Gaussian-noise simple linear regression model, over ways of estimating its parameters and some of the properties of the model, and over how to check the model's assumptions, we are now ready to begin doing

some serious statistical inference within the model[1]. In previous lectures, we came up with **point estimators** of the parameters and the conditional mean (prediction) function, but we weren't able to say much about the margin of uncertainty around these estimates. In this lecture we will focus on supplementing point estimates with *reliable* measures of uncertainty. This will naturally lead us to testing hypotheses about the true parameters — again, we will want hypothesis tests which are unlikely to get the answer wrong, whatever the truth might be.

To accomplish all this, we first need to understand the sampling distribution of our point estimators. We can find them, mathematically, but they involve the unknown true parameters in inconvenient ways. We will therefore work to find combinations of our estimators and the true parameters with fixed, parameter-free distributions; we'll get our confidence sets and our hypothesis tests from them.

Throughout this lecture, I am assuming, unless otherwise noted, that all of the assumptions of the Gaussian-noise simple linear regression model hold. After all, we checked those assumptions last time...

1 Sampling Distribution of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}^2$

The Gaussian-noise simple linear regression model has three parameters: the intercept β_0, the slope β_1, and the noise variance σ^2. We've seen, previously, how to estimate all of these by maximum likelihood; the MLE for the βs is the same as their least-squares estimates. These are

$$\hat{\beta}_1 = \frac{c_{XY}}{s_X^2} = \sum_{i=1}^{n} \frac{x_i - \overline{x}}{n s_X^2} y_i \tag{1}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \tag{2}$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 \tag{3}$$

We have also seen how to re-write the first two of these as a deterministic part plus a weighted sum of the noise terms ϵ:

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} \frac{x_i - \overline{x}}{n s_X^2} \epsilon_i \tag{4}$$

$$\hat{\beta}_0 = \beta_0 + \frac{1}{n} \sum_{i=1}^{n} \left(1 - \overline{x} \frac{x_i - \overline{x}}{s_X^2}\right) \epsilon_i \tag{5}$$

Finally, we have our modeling assumption that the ϵ_i are independent Gaussians, $\epsilon_i \sim N(0, \sigma^2)$.

[1] Presuming, of course, that the model's assumptions, when thoroughly checked, do in fact hold good.

1.1 Reminders of Basic Properties of Gaussian Distributions

Suppose $U \sim N(\mu, \sigma^2)$. By the basic algebra of expectations and variances, $\mathbb{E}[a + bU] = a + b\mu$, while $\mathrm{Var}[a + bU] = b^2\sigma^2$. This would be true of any random variable; a special property of Gaussians[2] is that $a + bU \sim N(a + b\mu, b^2\sigma^2)$.

Suppose $U_1, U_2, \ldots U_n$ are *independent* Gaussians, with means μ_i and variances σ_i^2. Then

$$\sum_{i=1}^{n} U_i \sim N(\sum_i \mu_i, \sum_i \sigma_i^2)$$

That the expected values add up for a sum is true of all random variables; that the variances add up is true for all uncorrelated random variables. That the sum follows the same type of distribution as the summands is a special property of Gaussians[3].

1.2 Sampling Distribution of $\hat{\beta}_1$

Since we're assuming Gaussian noise, the ϵ_i are independent Gaussians, $\epsilon_i \sim N(0, \sigma^2)$. Hence (using the first basic property of Gaussians)

$$\frac{x_i - \overline{x}}{n s_X^2}\epsilon_i \sim N(0, \left(\frac{x_i - \overline{x}}{n s_X^2}\right)^2 \sigma^2)$$

Thus, using the second basic property of Gaussians,

$$\sum_{i=1}^{n} \frac{x_i - \overline{x}}{n s_X^2}\epsilon_i \sim N(0, \sigma^2 \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{n s_X^2}\right)^2) \quad (6)$$

$$= N(0, \frac{\sigma^2}{n s_X^2}) \quad (7)$$

Using the first property of Gaussians again,

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{n s_X^2}) \quad (8)$$

This is the distribution of estimates we'd see if we repeated the experiment (survey, observation, etc.) many times, and collected the results. Every particular run of the experiment would give a slightly different $\hat{\beta}_1$, but they'd average out to β_1, the average squared difference from β_1 would be $\sigma^2/n s_X^2$, and a histogram of them would follow the Gaussian probability density function (Figure 2).

[2]There some other families of distributions which have this property; they're called "location-scale" families.

[3]There are some other families of distributions which have this property; they're called "stable" families.

```
# Simulate a Gaussian-noise simple linear regression model
# Inputs: x sequence; intercept; slope; noise variance; switch for whether to
  # return the simulated values, or run a regression and return the coefficients
# Output: data frame or coefficient vector
sim.gnslrm <- function(x, intercept, slope, sigma.sq, coefficients=TRUE) {
  n <- length(x)
  y <- intercept + slope*x + rnorm(n,mean=0,sd=sqrt(sigma.sq))
  if (coefficients) {
    return(coefficients(lm(y~x)))
  } else {
    return(data.frame(x=x, y=y))
  }
}

# Fix an arbitrary vector of x's
x <- seq(from=-5, to=5, length.out=42)
```

FIGURE 1: *Code setting up a simulation of a Gaussian-noise simple linear regression model, along a fixed vector of x_i values.*

It is a bit hard to use Eq. 8, because it involves two of the unknown parameters. We can manipulate it a bit to remove one of the parameters from the probability distribution,

$$\hat{\beta}_1 - \beta_1 \sim N(0, \frac{\sigma^2}{n s_X^2})$$

but that still has σ^2 on the right hand side, so we can't actually calculate anything. We could write

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma^2/\sqrt{n s_X^2}} \sim N(0,1)$$

but now we've got two unknown parameters on the left-hand side, which is also awkward.

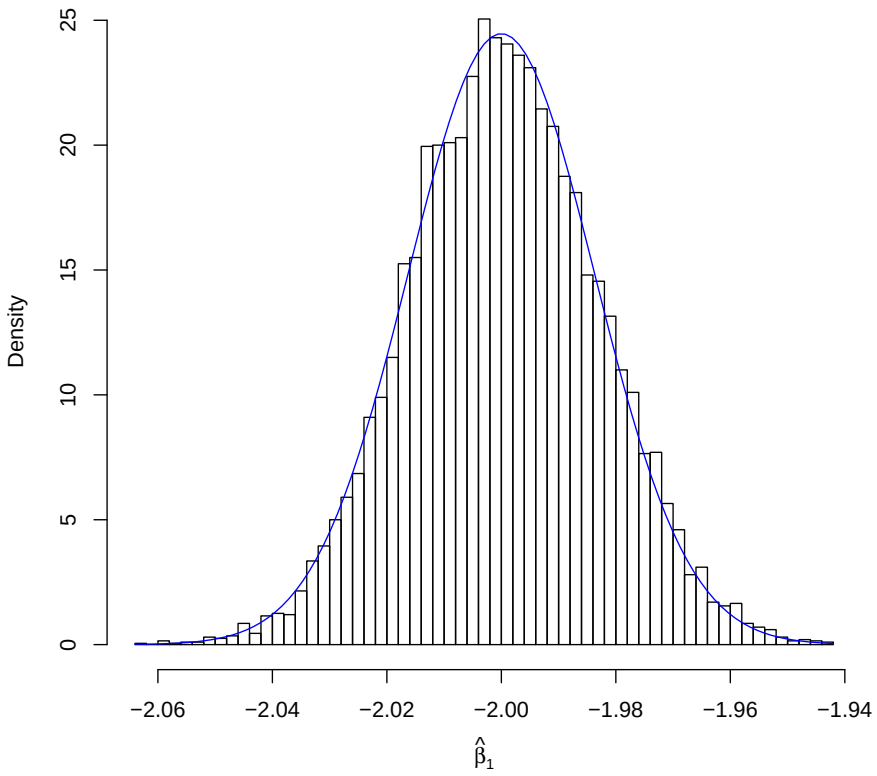
1.3 Sampling Distribution of $\hat{\beta}_0$

Starting from Eq. 5 rather than Eq. 4, an argument exactly parallel to the one we just went through gives

$$\hat{\beta}_0 \sim N(\beta_0, \frac{\sigma^2}{n}\left(1 + \frac{\overline{x}^2}{s_X^2}\right))$$

It follows, again by parallel reasoning, that

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\frac{\sigma^2}{n}\left(1 + \frac{\overline{x}^2}{s_X^2}\right)}} \sim N(0,1)$$




```
# Run the simulation 10,000 times and collect all the coefficients
  # What intercept, slope and noise variance does this impose?
many.coefs <- replicate(1e4, sim.gnslrm(x=x, 5, -2, 0.1, coefficients=TRUE))
# Histogram of the slope estimates
hist(many.coefs[2,], breaks=50, freq=FALSE, xlab=expression(hat(beta)[1]),
     main="")
# Theoretical Gaussian sampling distribution
theoretical.se <- sqrt(0.1/(length(x)*var(x)))
curve(dnorm(x,mean=-2,sd=theoretical.se), add=TRUE,
      col="blue")
```

FIGURE 2: *Simulating 10,000 runs of a Gaussian-noise simple linear regression model, calculating $\hat{\beta}_1$ each time, and comparing the histogram of estimates to the theoretical Gaussian distribution (Eq. 8, in blue).*

The right-hand side of this equation is admirably simple and easy for us to calculate, but the left-hand side unfortunately involves two unknown parameters, and that complicates any attempt to use it.

1.4 Sampling Distribution of $\hat{\sigma}^2$

It is mildly challenging, but certainly not too hard, to show that

$$\mathbb{E}\left[\hat{\sigma}^2\right] = \frac{n-2}{n}\sigma^2$$

As I have said before, this will be a problem on a future assignment, so I will not give a proof, but I will note that the way to proceed is to write

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 \ ;$$

then to write each residual e_i as a weighted sum of the noise terms ϵ; to use $\mathbb{E}\left[e_i^2\right] = \mathrm{Var}\left[e_i\right] + (\mathbb{E}\left[e_i\right])^2$; and finally to sum up over i.

Notice that this implies that $\mathbb{E}\left[\hat{\sigma}^2\right] = 0$ when $n = 2$. This is because any two points in the plane define a (unique) line, so if we have only two data points, least squares will just run a line through them exactly, and have an in-sample MSE of 0. In general, we get the factor of $n - 2$ from the fact that we are estimating two parameters.

We can however be much more specific. When $\epsilon_i \sim N(0, \sigma^2)$, it can be shown that

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$$

Notice, by the way, that this equation involves no unknown parameters on the right-hand side, and only one on the left-hand side. It lets us calculate the probability that $\hat{\sigma}^2$ is within any given *factor* of σ^2. If, for instance, we wanted to know the probability that $\hat{\sigma}^2 \geq 7\sigma^2$, this will let us find it.

I will offer only a hand-waving explanation; I am afraid I am not aware of any truly elementary mathematical explanation — every one I know of either uses probability facts which are about as obscure as the result to be shown, or linear-algebraic facts about the properties of idempotent matrices[4], and we've not seen, *yet*, how to write linear regression in matrix form. I do however want to re-assure you that there are actual proofs, and I promise to include one in these notes once we've seen how to connect what we're doing to matrices and linear algebra.

I am afraid I do not have even a hand-waving explanation of a second important property of $\hat{\sigma}^2$: it is statistically independent of $\hat{\beta}_0$ and $\hat{\beta}_1$. This is *not* obvious — after all, all three of these estimators are functions of the same noise variables ϵ — but it *is* true, and, again, I promise to provide a genuine proof in these notes once we've gone over the necessary math.

[4] Where $M^2 = M$.

1.4.1 The Hand-Waving Explanation for $n-2$

Let's think for a moment about a related (but strictly different!) quantity from $\hat{\sigma}^2$, namely

$$\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2$$

This is a weighted sum of independent, mean-zero squared Gaussians, which is where the connection to χ^2 distributions comes in.

Some reminders about χ^2 If $Z \sim N(0,1)$, then $Z^2 \sim \chi^2_1$ *by definition* (of the χ^2_1 distribution). From this, it follows that $\mathbb{E}\left[\chi^2_1\right] = 1$, $\mathrm{Var}\left[\chi^2_1\right] = \mathbb{E}\left[Z^4\right] - (\mathbb{E}\left[Z^2\right])^2 = 2$. If $Z_1, Z_2, \ldots Z_d \sim N(0,1)$ and are independent, then the χ^2_d distribution is *defined* to be the distribution of $\sum_{i=1}^{d} Z_i^2$. By simple algebra, it follows that $\mathbb{E}\left[\chi^2_d\right] = d$ while $\mathrm{Var}\left[\chi^2_d\right] = 2d$.

Back to the sum of squared noise terms ϵ_i isn't a standard Gaussian, but ϵ_i/σ is, so

$$\frac{\sum_{i=1}^{n}\epsilon_i^2}{\sigma^2} = \sum_{i=1}^{n}(\frac{\epsilon_i}{\sigma})^2 \sim \chi^2_n$$

The numerator here is *like* $n\hat{\sigma}^2 = \sum_i e_i^2$, but of course the residuals e_i are not the same as the noise terms ϵ_i.

The reason we end up with a χ^2_{n-2} distribution, rather than a χ^2_n distribution, is that we're estimating two parameters from the data removes two degrees of freedom, so two of the ϵ_i end up making no real contribution to the sum of squared errors. (Again, if $n = 2$, we'd be able to fit the two data points *exactly* with the least squares line.) If we had estimated more or fewer parameters, we would have had to adjust the number of degrees of freedom accordingly.

(There is also a geometric interpretation: the sum of squared errors, $\sum_{i=1}^{n} e_i^2$, is the squared length of the n-dimensional vector of residuals, $(e_1, e_2, \ldots e_n)$. But the residuals must obey the two equations $\sum_i e_i = 0$, $\sum_i x_i e_i = 0$, so the residual vector actually is confined to an $(n-2)$-dimensional linear subspace. Thus we only end up adding up $(n-2)$ *independent* contributions to its length. If we estimated more parameters, we'd have more estimating equations, and so more constraints on the vector of residuals.)

1.5 Standard Errors of $\hat{\beta}_0$ and $\hat{\beta}_1$

The **standard error** of an estimator is its standard deviation[5]. We've just seen that the true standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$ are, respectively,

$$\text{se}\left[\hat{\beta}_1\right] = \frac{\sigma}{s_x\sqrt{n}} \tag{9}$$

$$\text{se}\left[\hat{\beta}_0\right] = \frac{\sigma}{\sqrt{n}s_X}\sqrt{s_X^2 + \overline{x}^2} \tag{10}$$

Unfortunately, these standard errors involve the unknown parameter σ^2 (or its square root σ, equally unknown to us).

We can, however, *estimate* the standard errors. The maximum-likelihood estimates just substitute $\hat{\sigma}$ for σ:

$$\widehat{\text{se}}\left[\hat{\beta}_1\right] = \frac{\hat{\sigma}}{s_x\sqrt{n}} \tag{11}$$

$$\widehat{\text{se}}\left[\hat{\beta}_0\right] = \frac{\hat{\sigma}}{s_X\sqrt{n}}\sqrt{s_X^2 + \overline{x}^2} \tag{12}$$

For later theoretical purposes, however, things will work out slightly nicer if we use the de-biased version, $\frac{n}{n-2}\hat{\sigma}^2$:

$$\widehat{\text{se}}\left[\hat{\beta}_1\right] = \frac{\hat{\sigma}}{s_x\sqrt{n-2}} \tag{13}$$

$$\widehat{\text{se}}\left[\hat{\beta}_0\right] = \frac{\hat{\sigma}}{s_x\sqrt{n-2}}\sqrt{s_X^2 + \overline{x}^2} \tag{14}$$

These standard errors — approximate or estimated though they be — are one important way of quantifying how much uncertainty there is around our point estimates. However, we can't use them, *alone* to say anything terribly precise[6] about, say, the probability that β_1 is in the interval $[\hat{\beta}_1 - \widehat{\text{se}}\left[\hat{\beta}_1\right], \hat{\beta}_1 - \widehat{\text{se}}\left[\hat{\beta}_1\right]]$, which is the sort of thing we'd want to be able to give guarantees about the reliability of our estimates.

2 Sampling distribution of $(\hat{\beta} - \beta)/\widehat{\text{se}}\left[\hat{\beta}\right]$

It should take only a little work with the properties of the Gaussian distribution to convince yourself that

$$\frac{\hat{\beta}_1 - \beta_1}{\text{se}\left[\hat{\beta}_1\right]} \sim N(0,1)$$

[5] We don't just call it the standard deviation because we want to emphasize that it is, in fact, telling us about the random errors our estimator makes.

[6] Exercise to think through: Could you use Chebyshev's inequality (the extra credit problem from Homework 1) here?

the standard Gaussian distribution. If the Oracle told us σ^2, we'd know $\mathrm{se}\left[\hat{\beta}_1\right]$, and so we could assert that (for example)

$$\begin{aligned}
\mathbb{P}\left(\beta_1 - 1.96\mathrm{se}\left[\hat{\beta}_1\right] \leq \hat{\beta}_1 \leq \beta_1 + 1.96\mathrm{se}\left[\hat{\beta}_1\right]\right) & (15)\\
= \mathbb{P}\left(-1.96\mathrm{se}\left[\hat{\beta}_1\right] \leq \hat{\beta}_1 - \beta_1 \leq 1.96\mathrm{se}\left[\hat{\beta}_1\right]\right) & (16)\\
= \mathbb{P}\left(-1.96 \leq \frac{\hat{\beta}_1 - \beta_1}{\mathrm{se}\left[\hat{\beta}_1\right]} \leq 1.96\right) & (17)\\
= \Phi(1.96) - \Phi(-1.96) = 0.95 & (18)
\end{aligned}$$

where Φ is the cumulative distribution function of the $N(0,1)$ distribution.

Since the oracles have fallen silent, we can't use this approach. What we *can* do is use the following fact[7]:

Proposition 1 *If $Z \sim N(0,1)$, $S^2 \sim \chi^2_d$, and Z and S^2 are independent, then*

$$\frac{Z}{\sqrt{S^2/d}} \sim t_d$$

(I call this a proposition, but it's almost a definition of what we mean by a t distribution with d degrees of freedom. Of course, if we take this as the definition, the proposition that this distribution has a probability density $\propto (1 + x^2/d)^{-(d+1)/2}$ would become yet another proposition to be demonstrated.)

Let's try to manipulate $(\hat{\beta}_1 - \beta_1)/\widehat{\mathrm{se}}\left[\hat{\beta}_1\right]$ into this form.

[7] When I messed up the derivation in class today, I left out dividing by d in the denominator. As I mentioned at the end of that debacle, this was stupid. As $d \to \infty$, t_d converges on the standard Gaussian distribution $N(0,1)$. (Notice that $\mathbb{E}\left[d^{-1}\chi^2_d\right] = 1$, while $\mathrm{Var}\left[d^{-1}\chi^2_d\right] = 2/d$, so $d^{-1}\chi^2_d \to 1$.) Without the normalizing factor of d inside the square root, however, looking just at Z/S, we've got a random variable whose distribution doesn't change with d being divided by something whose magnitude *grows* with d, so $Z/S \to 0$ as $d \to \infty$, not $\to N(0,1)$. I apologize again for my error.

$$\frac{\hat{\beta}_1 - \beta_1}{\widehat{\text{se}}\left[\hat{\beta}_1\right]} = \frac{\hat{\beta}_1 - \beta_1}{\sigma} \frac{\sigma}{\widehat{\text{se}}\left[\hat{\beta}_1\right]} \tag{19}$$

$$= \frac{\frac{\hat{\beta}_1 - \beta_1}{\sigma}}{\frac{\widehat{\text{se}}\left[\hat{\beta}_1\right]}{\sigma}} \tag{20}$$

$$= \frac{N(0, 1/ns_X^2)}{\frac{\hat{\sigma}}{s_x \sigma\sqrt{n-2}}} \tag{21}$$

$$= \frac{s_X N(0, 1/ns_X^2)}{\frac{\hat{\sigma}}{\sigma\sqrt{n-2}}} \tag{22}$$

$$= \frac{N(0, 1/n)}{\frac{\hat{\sigma}}{\sigma\sqrt{n-2}}} \tag{23}$$

$$= \frac{\sqrt{n}N(0, 1/n)}{\frac{\sqrt{n}\hat{\sigma}}{\sigma\sqrt{n-2}}} \tag{24}$$

$$= \frac{N(0,1)}{\sqrt{\frac{n\hat{\sigma}^2}{\sigma^2}\frac{1}{n-2}}} \tag{25}$$

$$= \frac{N(0,1)}{\sqrt{\chi^2_{n-2}/(n-2)}} \tag{26}$$

$$= t_{n-2} \tag{27}$$

where in the last step I've used the proposition I stated (without proof) above.

To sum up:

Proposition 2 *Using the* $\widehat{\text{se}}\left[\hat{\beta}_1\right]$ *of Eq. 13,*

$$\frac{\hat{\beta}_1 - \beta_1}{\widehat{\text{se}}\left[\hat{\beta}_1\right]} \sim t_{n-2} \tag{28}$$

Notice that we can compute $\widehat{\text{se}}\left[\hat{\beta}_1\right]$ without knowing any of the true parameters — it's a pure statistic, just a function of the data. This is a key to actually using the proposition for anything useful.

By exactly parallel reasoning, we may also demonstrate that

$$\frac{\hat{\beta}_0 - \beta_0}{\widehat{\text{se}}\left[\hat{\beta}_0\right]} \sim t_{n-2}$$

3 Sampling Intervals for $\hat{\beta}$; hypothesis tests for $\hat{\beta}$

Let's trace through one of the consequences of Eq. 28. For any $k > 0$,

$$\mathbb{P}\left(\beta_1 - k\widehat{\text{se}}\left[\hat{\beta}_1\right] \leq \hat{\beta}_1 \leq \beta_1 + k\widehat{\text{se}}\left[\hat{\beta}_1\right]\right) \tag{29}$$

$$= \mathbb{P}\left(k\widehat{\text{se}}\left[\hat{\beta}_1\right] \leq \hat{\beta}_1 - \beta_1 \leq k\widehat{\text{se}}\left[\hat{\beta}_1\right]\right) \tag{30}$$

$$= \mathbb{P}\left(k \leq \frac{\hat{\beta}_1 - \beta_1}{\widehat{\text{se}}\left[\hat{\beta}_1\right]} \leq k\right) \tag{31}$$

$$= \int_{-k}^{k} t_{n-2}(u)du \tag{32}$$

where by a slight abuse of notation I am writing $t_{n-2}(u)$ for the probability density of the t distribution with $n-2$ degrees of freedom, evaluated at the point u.

It should be evident that if you pick any α between 0 and 1, I can find a $k(n, \alpha)$ such that

$$\int_{-k(n,\alpha)}^{k(n,\alpha)} t_{n-2}(u)du = 1 - \alpha$$

I therefore define the (symmetric) $1-\alpha$ **sampling interval** for $\hat{\beta}_1$, when the true slope is β_1, as

$$[\beta_1 - k(n,\alpha)\widehat{\text{se}}\left[\hat{\beta}_1\right], \beta_1 + k(n,\alpha)\widehat{\text{se}}\left[\hat{\beta}_1\right]] \tag{33}$$

If the true slope is β_1, then $\hat{\beta}_1$ will be within this sampling interval with probability $1-\alpha$. This leads directly to a test of the null hypothesis that the slope $\beta_1 = \beta_1^*$: reject the null if $\hat{\beta}_1$ is outside the sampling interval for β_1^*, and retain the null when $\hat{\beta}_1$ is inside that sampling interval. This test is called the **Wald test**, after the great statistician Abraham Wald[8].

By construction, the Wald test's probability of rejection under the null hypothesis — the **size**, or **type I error rate**, or **false alarm rate** of the test — is exactly α. Of course, the other important property of a hypothesis test is its **power** — the probability of rejecting the null when it is false. From Eqn. 28, it should be clear that if the true $\beta_1 \neq \beta_1^*$, the probability that $\hat{\beta}_1$ is inside the sampling interval for β_1^* is $< 1-\alpha$, with the difference growing as $|\beta_1 - \beta_1^*|$ grows. An exact calculation could be done (it'd involve what's called the "non-central t distribution"), but is not especially informative. The point is that the power is always $> \alpha$, and grows with the departure from the null hypothesis.

[8] As is common with eponyms in the sciences, Wald was not, in fact, the first person to use the test, but he made one of the most important early studies of its properties, and he was already famous for other reasons.

If you were an economist, psychologist, or something of their ilk, you have a powerful drive — almost a spinal reflex not involving the higher brain regions — to test whether $\beta_1 = 0$. Under the Wald test, you would reject that point null hypothesis when $|\hat{\beta}_1|$ exceeds a certain number of standard deviations. As an intelligent statistician in control of your own actions, you would read the section on "statistical significance" below, before doing any such thing.

All of the above applies, *mutatis mutandis*, to $\frac{\hat{\beta}_0 - \beta_0}{\widehat{\text{se}}\left[\hat{\beta}_0\right]}$.

4 Building Confidence Intervals from Sampling Intervals

Once we know how to calculate sampling intervals, we can plot the sampling interval for every possible value of β_1 (Figure 3). They're the region marked off by two parallel lines, one $k(n,\alpha)\widehat{\text{se}}\left[\hat{\beta}_1\right]$ above the main diagonal and one equally far below the main diagonal.

The sampling intervals (as in Figure 3) are theoretical constructs — mathematical consequences of the assumptions in the the probability model that (we hope) describes the world. After we gather data, we can actually calculate $\hat{\beta}_1$. This is a random quantity, but it will have some particular value on any data set. We can mark this realized value, and draw a horizontal line across the graph at that height (Figure 4).

The $\hat{\beta}_1$ we observed is within the sampling interval for some (possible or hypothetical) values of β_1, and outside the sampling interval for others. We define the **confidence set**, with **confidence level** $1 - \alpha$, as

$$\left\{\beta_1 \ : \ \hat{\beta}_1 \in [\beta_1 - k(n,\alpha)\widehat{\text{se}}\left[\hat{\beta}_1\right], \beta_1 + k(n,\alpha)\widehat{\text{se}}\left[\hat{\beta}_1\right]]\right\} \tag{34}$$

This is precisely the set of β_1 which we retain when we run the Wald test with size α. In other words: we test every possible β_1; if we'd reject that null hypothesis, that value of β_1 gets removed from the hypothesis test; if we'd retain that null, β_1 stays in the confidence set[9]. Figure 5 illustrate a confidence set, and shows (unsurprisingly) that in this case the confidence set is indeed a confidence *interval*. Indeed, a little manipulation of Eq. 34 gives us an explicit formula for the confidence set, which is an interval:

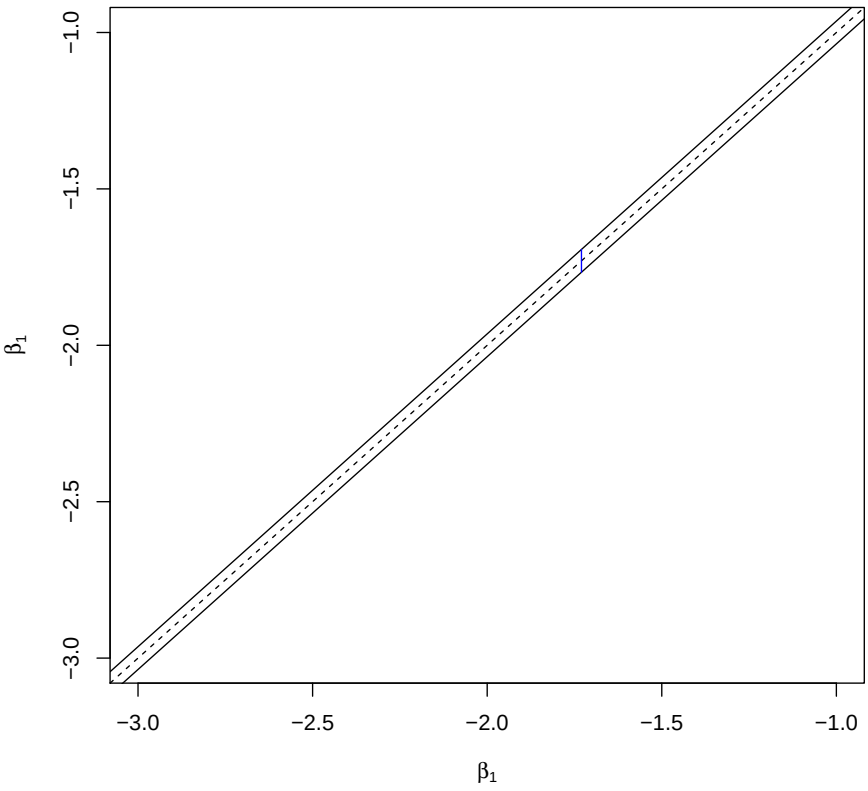
Confidence set = Test all the hypotheses!

$$[\hat{\beta}_1 - k(n,\alpha)\widehat{\text{se}}\left[\hat{\beta}_1\right], \hat{\beta}_1 + k(n,\alpha)\widehat{\text{se}}\left[\hat{\beta}_1\right]$$

The correct interpretation of a confidence set is that it offers us a dilemma. One of two[10] things must be true:

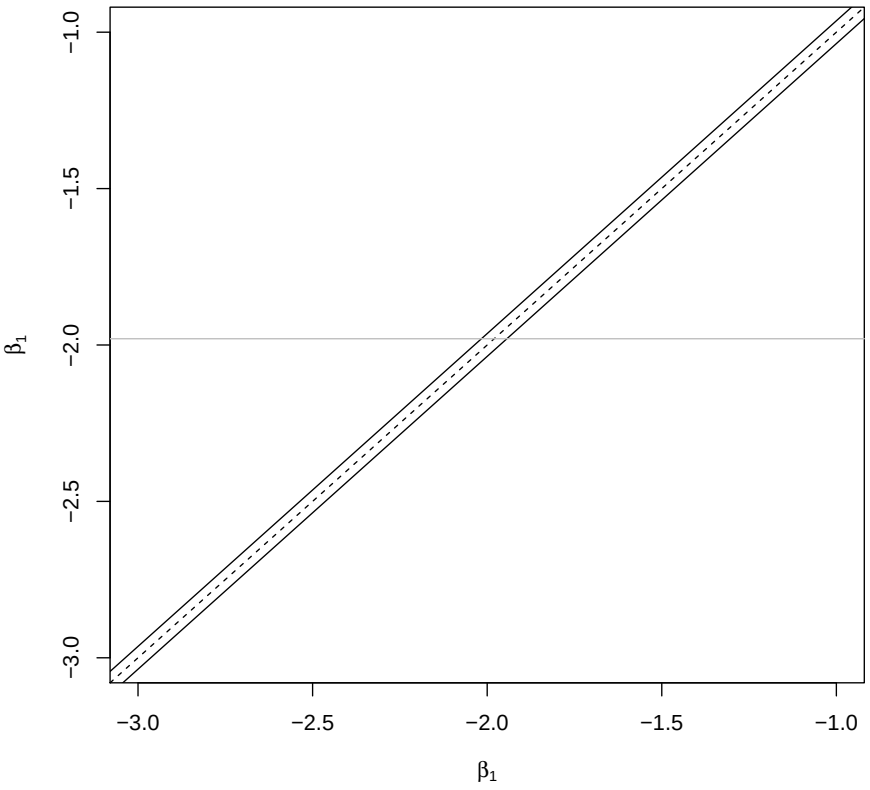
[9] Cf. the famous Sherlock Holmes line "When you have eliminated the impossible, whatever remains, however improbable, must be the truth." In forming the confidence set, we are eliminating the merely *unlikely*, rather than the absolutely impossible. This is because, not living in a detective story, we get only noisy and imperfect evidence.

[10] Strictly speaking, there is a third option: our model could be wrong. Hence the importance of model checking *before* doing within-model inference.



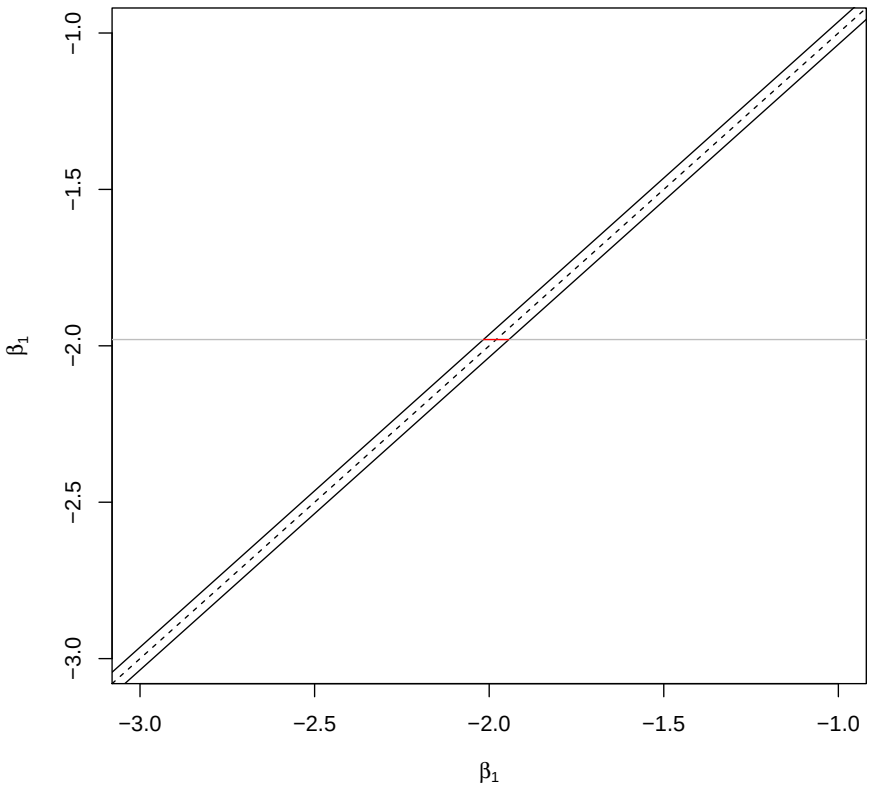

```
lm.sim <- lm(y~x, data=sim.gnslrm(x=x, 5, -2, 0.1, coefficients=FALSE))
hat.sigma.sq <- mean(residuals(lm.sim)^2)
se.hat.beta.1 <- sqrt(hat.sigma.sq/(var(x)*(length(x)-2)))
alpha <- 0.02
k <- qt(1-alpha/2, df=length(x)-2)
plot(0, xlim=c(-3,-1),ylim=c(-3,-1),type="n",
     xlab=expression(beta[1]),
     ylab=expression(hat(beta)[1]), main="")
abline(a=k*se.hat.beta.1,b=1)
abline(a=-k*se.hat.beta.1,b=1)
abline(a=0,b=1,lty="dashed")
beta.1.star <- -1.73
segments(x0=beta.1.star,y0=k*se.hat.beta.1+beta.1.star,
         x1=beta.1.star,y1=-k*se.hat.beta.1+beta.1.star,
         col="blue")
```

FIGURE 3: *Sampling intervals for $\hat{\beta}_1$ as a function of β_1. For compatibility with the earlier simulation, I have set $n = 42$, $s_X^2 = 9$, and (from one run of the model) $\hat{\sigma}^2 = 0.081$; and, just because $\alpha = 0.05$ is cliched, $\alpha = 0.02$. Equally arbitrarily, the blue vertical line illustrates the sampling interval when $\beta_1 = -1.73$.*




```
plot(0, xlim=c(-3,-1),ylim=c(-3,-1),type="n",
     xlab=expression(beta[1]),
     ylab=expression(hat(beta)[1]), main="")
abline(a=k*se.hat.beta.1,b=1)
abline(a=-k*se.hat.beta.1,b=1)
abline(a=0,b=1,lty="dashed")
beta.1.hat <- coefficients(lm.sim)[2]
abline(h=beta.1.hat,col="grey")
```

FIGURE 4: *As in Figure 3, but with the addition of a horizontal line marking the observed value of $\hat{\beta}_1$ on a particular realization of the simulation (in grey).*




```
plot(0, xlim=c(-3,-1),ylim=c(-3,-1),type="n",
     xlab=expression(beta[1]),
     ylab=expression(hat(beta)[1]), main="")
abline(a=k*se.hat.beta.1,b=1)
abline(a=-k*se.hat.beta.1,b=1)
abline(a=0,b=1,lty="dashed")
beta.1.hat <- coefficients(lm.sim)[2]
abline(h=beta.1.hat,col="grey")
segments(x0=beta.1.hat-k*se.hat.beta.1, y0=beta.1.hat,
         x1=beta.1.hat+k*se.hat.beta.1, y1=beta.1.hat,
         col="red")
```

FIGURE 5: *As in Figure 4, but with the* **confidence set** *marked in red. This is the collection of all* β_1 *where* $\hat{\beta}_1$ *falls within the* $1 - \alpha$ *sampling interval.*

1. The true β_1 is inside the confidence set.
2. $\hat{\beta}_1$ is outside the sampling interval of the true β_1.

We know that the second option has probability at most α, no matter what the true β_1 is, so we may rephrase the dilemma. Either

1. The true β_1 is inside the confidence set, or
2. We're very unlucky, because something whose probability is $\leq \alpha$ happened.

Since, most of the time, we're not very unlucky, the confidence set is, in fact, a reliable way of giving a margin of error for the true parameter β_1.

Width of the confidence interval Notice that the width of the confidence interval is $2k(n,\alpha)\widehat{\text{se}}\left[\hat{\beta}_1\right]$. This tells us what controls the width of the confidence interval:

1. As α shrinks, the interval widens. (High confidence comes at the price of big margins of error.)
2. As n grows, the interval shrinks. (Large samples mean precise estimates.)
3. As σ^2 increases, the interval widens. (The more noise there is around the regression line, the less precisely we can measure the line.)
4. As s_X^2 grows, the interval shrinks. (Widely-spread measurements give us a precise estimate of the slope.)

What about β_0? By exactly parallel reasoning, a $1-\alpha$ confidence interval for β_0 is $[\hat{\beta}_0 - k(n,\alpha)\widehat{\text{se}}\left[\hat{\beta}_0\right], \hat{\beta}_0 + k(n,\alpha)\widehat{\text{se}}\left[\hat{\beta}_0\right]]$.

What about σ^2? See Exercise 1.

What α should we use? It's become conventional to set $\alpha = 0.05$. To be honest, this owes more to the fact that the resulting k tends to 1.96 as $n \to \infty$, and $1.96 \approx 2$, and most psychologists and economists could multiply by 2, even in 1950, than to any genuine principle of statistics or scientific method. A 5% error rate corresponds to messing up about one working day in every month, which you might well find high. On the other hand, there is nothing which stops you from increasing α. It's often illuminating to plot a series of confidence sets, at different values of α.

What about power? The **coverage** of a confidence set is the probability that it includes the true parameter value. This is not, however, the only virtue we want in a confidence set; if it was, we could just say "Every possible parameter is in the set", and have 100% coverage no matter what. We would also like the *wrong* values of the parameter to have a high probability of *not* being in the set. Just as the coverage is controlled by the size / false-alarm probability / type-I error rate α of the hypothesis test, the probability of excluding the wrong parameters is controlled by the power / miss probability / type-II error rate. Test with higher power exclude (correctly) more parameter values, and give smaller confidence sets.

4.1 Confidence Sets and Hypothesis Tests

I have derived confidence sets for β by inverting a specific hypothesis test, the Wald test. There is a more general relationship between confidence sets and hypothesis tests.

1. Inverting any hypothesis test gives us a confidence set.
2. If we have a way of constructing a $1-\alpha$ confidence set, we can use it to test the hypothesis that $\beta = \beta^*$: reject when β^* is outside the confidence set, retain the null when β^* is inside the set.

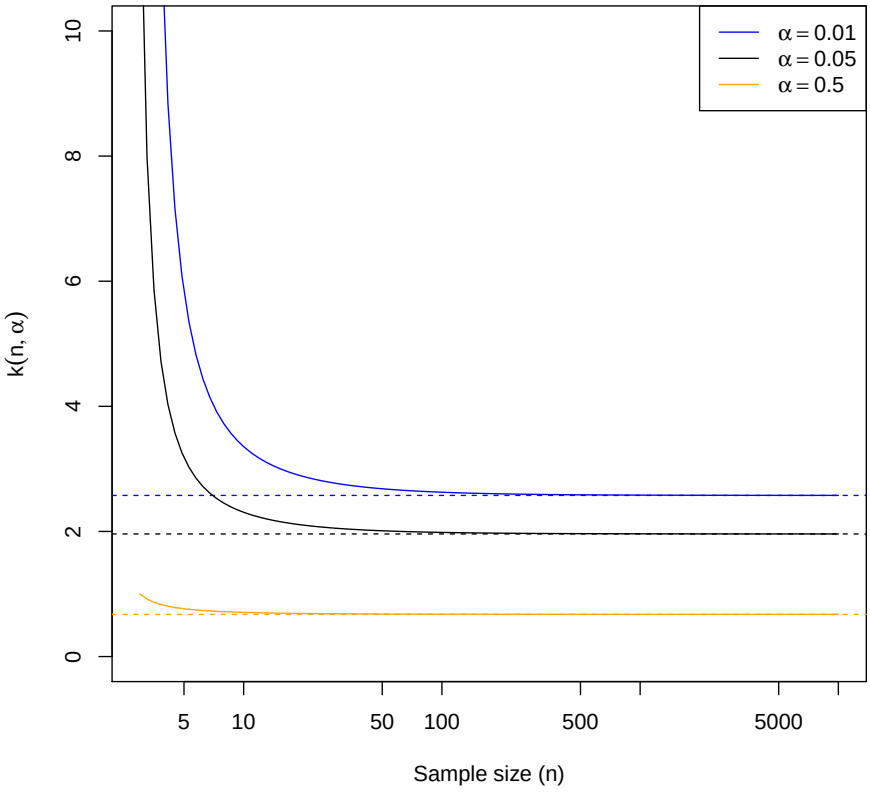
I will leave it as a pair of exercises (2 and 3) to that inverting a test of size α gives a $1-\alpha$ confidence set, and that inverting a $1-\alpha$ confidence set gives a test of size α.

4.2 Large-n Asymptotics

As $n \to \infty$, $\hat{\sigma}^2 \to \sigma^2$. It follows (by continuity) that $\widehat{\text{se}}\left[\hat{\beta}\right] \to \text{se}\left[\hat{\beta}\right]$. Hence,

$$\frac{\hat{\beta} - \beta}{\widehat{\text{se}}\left[\hat{\beta}\right]} \to N(0,1)$$

which considerably simplifies the sampling intervals and confidence sets; as n grows, we can forget about the t distribution and just use the standard Gaussian distribution. Figure 6 plots the convergence of $k(n,\alpha)$ towards the $k(\infty,\alpha)$ we'd get from the Gaussian approximation. As you can see from the figure, by the time $n = 100$ —a quite small data set by modern standards — the difference between the t distribution and the standard-Gaussian is pretty trivial.




```
curve(qt(0.995,df=x-2),from=3,to=1e4,log="x", ylim=c(0,10),
      xlab="Sample size (n)", ylab=expression(k(n,alpha)),col="blue")
abline(h=qnorm(0.995),lty="dashed",col="blue")
curve(qt(0.975,df=x-2), add=TRUE)
abline(h=qnorm(0.975),lty="dashed")
curve(qt(0.75,df=x-2), add=TRUE, col="orange")
abline(h=qnorm(0.75), lty="dashed", col="orange")
legend("topright", legend=c(expression(alpha==0.01), expression(alpha==0.05),
                            expression(alpha==0.5)),
       col=c("blue","black","orange"), lty="solid")
```

FIGURE 6: *Convergence of* $k(n, \alpha)$ *as* $n \to \infty$, *illustrated for* $\alpha = 0.01$, $\alpha = 0.05$ *and* $\alpha = 0.5$. *(Why do I plot the* 97.5^{th} *percentile when I'm interested in* $\alpha = 0.05$*?)*

5 Statistical Significance: Uses and Abuses

5.1 p-Values

The test statistic for the Wald test,

$$T = \frac{\hat{\beta}_1 - \beta_1^*}{\widehat{\text{se}}\left[\hat{\beta}_1\right]}$$

has the nice, intuitive property that it ought to be close to zero when the null hypothesis $\beta_1 = \beta_1^*$ is true, and take large values (either positive or negative) when the null hypothesis is false. When a test statistic works like this, it makes sense to summarize just how bad the data looks for the null hypothesis in a **p-value**: when our observed value of the test statistic is T_{obs}, the p-value is

$$P = \mathbb{P}\left(|T| \geq |T_{obs}|\right)$$

calculating the probability under the null hypothesis. (I write a capital P here as a reminder that this is a random quantity, though it's conventional to write the phrase "p-value" with a lower-case p.) This is the probability, under the null, of getting results which are at least as extreme as what we saw. It should be easy to convince yourself that rejecting the null in a level-α test is the same as getting a p-value $< \alpha$.

It is not too hard (Exercise 4) to show that P has a uniform distribution over $[0, 1]$ under the null hypothesis.

5.2 p-Values and Confidence Sets

When our test lets us calculate a p-value, we can form a $1 - \alpha$ confidence set by taking all the β's where the p-value is $\geq \alpha$. Conversely, if we have some way of making confidence sets already, we can get a p-value for the hypothesis $\beta = \beta^*$; it's the largest α such that β^* is in the $1 - \alpha$ confidence set.

5.3 Statistical Significance

If we test the hypothesis that $\beta_1 = \beta_1^*$ and reject it, we say that the difference between β_1 and β_1^* is **statistically significant**. Since, as I mentioned, many professions have an overwhelming urge to test the hypothesis $\beta_1 = 0$, it's common to hear people say that "β_1 is statistically significant" when they mean "β_1 is difference from 0 is statistically significant".

This is harmless enough, as long as we keep firmly in mind that "significant" is used here as a technical term, with a special meaning, and is *not* the same as "important", "relevant", etc. When we reject the hypothesis that $\beta_1 = 0$, what we're saying is "It's really implausibly hard to fit this data with a flat line, as opposed to one with a slope". This is informative, if we had serious reasons to think that a flat line was a live option.

It is incredibly common for researchers from other fields, and even some statisticians, to reason as follows: "I tested whether $\beta_1 = 0$ or not, and I retained the null; *therefore* β_1 is *in*significant, and I can ignore it." This is, of course, a complete fallacy.

To see why, it is enough to realize that there are (at least) two reasons why our hypothesis test might retain the null $\beta_1 = 0$:

1. β_1 is, in fact, zero,
2. $\beta_1 \neq 0$, but $\widehat{\text{se}}\left[\hat{\beta}_1\right]$ is so large that we can't tell anything about β_1 with any confidence.

There is a very big difference between data which lets us say "we can be quite confident that the true β_1 is, if not perhaps exactly 0, then very small", and data which only lets us say "we have no earthly idea what β_1 is, and it may as well be zero for all we can tell"[11]. It is good practice to always compute a confidence interval, but it is *especially* important to do so when you retain the null, so you know whether you can say "this parameter is zero to within such-and-such a (small) precision", or whether you have to admit "I couldn't begin to tell you what this parameter is".

Substantive vs. statistical significance Even a huge β_1, which it would be crazy to ignore in any circumstance, can be statistically insignificant, so long as $\widehat{\text{se}}\left[\hat{\beta}_1\right]$ is large enough. Conversely, any β_1 which isn't exactly zero, no matter how close it might be to 0, will become statistically significant at any threshold once $\widehat{\text{se}}\left[\hat{\beta}_1\right]$ is small enough. Since, as $n \to \infty$,

$$\widehat{\text{se}}\left[\hat{\beta}_1\right] \to \frac{\sigma}{s_X\sqrt{n}}$$

we can show that $\widehat{\text{se}}\left[\hat{\beta}_1\right] \to 0$, and $\frac{\hat{\beta}_1}{\widehat{\text{se}}\left[\hat{\beta}_1\right]} \to \pm\infty$, unless β_1 is exactly 0 (see below).

Statistical significance is a weird mixture of how big the coefficient is, how big a sample we've got, how much noise there is around the regression line, and how spread out the data is along the x axis. This has so little to do with "significance" in ordinary language that it's pretty unfortunate we're stuck with the word; if the Ancestors had decided to say "statistically detectable" or "statistically distinguishable from 0", we might have avoided a lot of confusion.

If *you* confuse substantive and statistical significance in this class, it will go badly for you.

[11] Imagine hearing what sounds like the noise of an animal in the next room. If the room is small, brightly lit, free of obstructions, and you make a thorough search of it with unimpaired vision and concentration, not finding an animal in it is, in fact, good evidence that there was no animal there to be found. If on the other hand the room is dark, large, full of hiding places, and you make a hurried search while distracted, without your contact lenses and after a few too many drinks, you could easily have missed all sorts of things, and your negative report has little weight as evidence. (In this parable, the difference between a large $|\beta_1|$ and a small $|\beta_1|$ is the difference between looking for a Siberian tiger and looking for a little black cat.)

5.4 Appropriate Uses of p-Values and Significance Testing

I do not want this section to give the impression that p-values, hypothesis testing, and statistical significance are unimportant or necessarily misguided. They're often used badly, but that's true of every statistical tool from the sample mean on down the line. There are certainly situations where we really do want to know whether we have good evidence against some *exact* statistical hypothesis, and that's just the job these tools do. What are some of these situations?

Model checking Our statistical models often make very strong, claims about the probability distribution of the data, with little wiggle room. The simple linear regression model, for instance, claims that the regression function is *exactly* linear, and that the noise around this line has *exactly* constant variance. If we test these claims and find very small p-values, then we have evidence that there's a detectable, systematic departure from the model assumptions, and we should re-formulate the model.

Actual scientific interest Some scientific theories make very precise predictions about coefficients. According to Newton, the gravitational force between two masses is inversely proportional to the *square* of the distance between them, $\propto r^{-2}$. The prediction is exactly $\propto r^{-2}$, not $\propto r^{-1.99}$ nor $\propto r^{-2.05}$. Measuring that exponent and finding even tiny departures from 2 would be big news, if we had reason to think they were real and not just noise[12]. One of the most successful theories in physics, quantum electrodynamics, makes predictions about some properties of hydrogen atoms with a theoretical precision of one part in a trillion; finding even tiny discrepancies between what the theory predicts and what we estimate would force us to rethink lots of physics[13]. Experiments to detect new particles, like the Higgs boson, essentially boil down to hypothesis testing, looking for deviations from theoretical predictions which should be exactly zero if the particle doesn't exist.

Outside of the natural sciences, however, it is harder to find examples of interesting, exact null hypothesis which are, so to speak, "live options". The best I can come up with are theories of economic growth and business cycles which predict that the share of national income going to labor (as opposed to capital) should be constant over time. Otherwise, in the social sciences, there's usually little theoretical reason to think that certain regression coefficients should be *exactly* zero, or *exactly* one, or anything else.

Neutral models A partial exception is the use of **neutral models**, which comes out of genetics and ecology. The idea here is to check whether some mechanism is at work in a particular situation — say, whether some gene is

[12] In fact, it *was* big news: Einstein's theory of general relativity.

[13] Feynman (1985) gives a great conceptual overview of quantum electrodynamics. Currently, theory agrees with experiment to the limits of experimental precision, which is only about one part in a billion (`https://en.wikipedia.org/wiki/Precision_tests_of_QED`).

subject to natural selection. One constructs two models; one incorporates all the mechanisms (which we think are) at work, including the one under investigation, and the other incorporate all the *other* mechanisms, but "neutralizes" the one of interest. (In a genetic example, the neutral model would probably incorporate the effects of mutation, sexual repdouction, the random sampling of which organisms become the ancestors of the next generation, perhaps migration, etc. The non-neutral model would include all this *plus* the effects of natural selection.) Rejecting the neutral model in favor of the non-neutral one then becomes evidence that the disputed mechanism is needed to explain the data.

In the cases where this strategy has been done well, the neutral model is usually a pretty sophisticated stochastic model, and the "neutralization" is not as simple as just setting some slope to zero. Nonetheless, this is a situation where we do actually learn something about the world by testing a null hypothesis.

6 Any Non-Zero Parameter Becomes Significant with Enough Information

(This section is optional, but strongly recommended.)

Let's look more close at what happens to the test statistic when $n \to \infty$, and so at what happens to the p-value. Throughout, we'll be testing the null hypothesis that $\beta_1 = 0$, since this is what people most often do, but the same reasoning applies to departures from any fixed value of the slope. (Everything carries over with straightforward changes to testing hypotheses about the intercept β_0, too.)

We know that $\hat{\beta}_1 \sim N(\beta_1, \sigma^2/ns_X^2)$. This means[14]

$$\begin{aligned} \hat{\beta}_1 &\sim \beta_1 + N(0, \sigma^2/ns_X^2) && (35)\\ &= \beta_1 + \frac{\sigma}{s_X\sqrt{n}} N(0,1) && (36)\\ &= \beta_1 + O(1/\sqrt{n}) && (37) \end{aligned}$$

where $O(f(n))$ is read "order-of $f(n)$", meaning that it's a term whose size grows like $f(n)$ as $n \to \infty$, and we don't want (or need) to keep track of the details. Similarly, since $n\hat{sigma}^2/\sigma^2 \sim \chi^2_{n-2}$, we have[15]

$$\begin{aligned} n\hat{\sigma}^2 &\sim \sigma^2 \chi^2_{n-2} && (38)\\ \hat{\sigma}^2 &\sim \sigma^2 \frac{\chi^2_{n-2}}{n} && (39) \end{aligned}$$

[14] If seeing something like $\frac{\sigma}{s_X\sqrt{n}} N(0,1)$, feel free to introduce random variables $Z_n \sim N(0,1)$ (though not necessarily independent ones), and modify the equations accordingly.

[15] Again, feel free to introduce the random variable Ξ_n, which just so happens to have a χ^2_{n-2} distribution.

Since $\mathbb{E}\left[\chi^2_{n-2}\right] = n-2$ and $\mathrm{Var}\left[\chi^2_{n-2}\right] = 2(n-2)$,

$$\mathbb{E}\left[\frac{\chi^2_{n-2}}{n}\right] = \frac{n-2}{n} \to 1 \tag{40}$$

$$\mathrm{Var}\left[\frac{\chi^2_{n-2}}{n}\right] = \frac{2(n-2)}{n^2} \to 0 \tag{41}$$

with both limits happening as $n \to \infty$. In fact $\mathrm{Var}\left[\frac{\chi^2_{n-2}}{n}\right] = O(1/n)$, so

$$\hat{\sigma}^2 = \sigma^2\left(1 + O(1/\sqrt{n})\right) \tag{42}$$

Taking the square root, and using the fact[16] that $(1+x)^a \approx 1+ax$ when $|x| \ll 1$,

$$\hat{\sigma} = \sigma\left(1 + O(1/\sqrt{n})\right) \tag{43}$$

Put this together to look at our test statistic:

$$\frac{\hat{\beta}_1}{\widehat{\mathrm{se}}\left[\hat{\beta}_1\right]} = \frac{\beta_1 + O(1/\sqrt{n})}{\frac{\sigma\left(1+O(1/\sqrt{n})\right)}{s_X\sqrt{n}}} \tag{44}$$

$$= \sqrt{n}\frac{\beta_1 + O(1/\sqrt{n})}{(\sigma/s_X)\left(1+O(1/\sqrt{n})\right)} \tag{45}$$

$$= \sqrt{n}\frac{\beta_1}{\sigma/s_X}\left(1+O(1/\sqrt{n})\right) \tag{46}$$

$$= \sqrt{n}\frac{\beta_1}{\sigma/s_X} + O(1) \tag{47}$$

In words: so long as the true $\beta_1 \neq 0$, the test statistic is going to go off to $\pm\infty$, and the rate at which it escapes towards infinity is going to be proportional to $\sqrt{n}$. When we compare this against the null distribution, which is $N(0,1)$, eventually we'll get arbitrarily small p-values.

We can actually compute what those p-values should be, by two bounds on the standard Gaussian distribution[17]:

$$\left(\frac{1}{x} - \frac{1}{x^3}\right)\frac{1}{\sqrt{2\pi}}e^{-x^2/2} < 1 - \Phi(x) < \frac{1}{x}\frac{1}{\sqrt{2\pi}}e^{-x^2/2} \tag{48}$$

Thus

$$P_n = \mathbb{P}\left(|Z| \geq \left|\frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{n}s_X}a\right|\right) \tag{49}$$

$$= 2\mathbb{P}\left(Z \geq \left|\frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{n}s_X}\right|\right) \tag{50}$$

$$\leq \frac{2}{\sqrt{2\pi}}\frac{e^{-\frac{1}{2}\frac{\hat{\beta}_1^2}{\hat{\sigma}^2/ns_X^2}}}{\left|\frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{n}s_X}\right|} \tag{51}$$

[16] From the binomial theorem, back in high school algebra.

[17] See Feller (1957), Chapter VII, §1, Lemma 2. For a brief proof online, see `http://www.johndcook.com/normalbounds.pdf`.

To clarify the behavior, let's take the logarithm and divide by n:

$$\begin{aligned}\frac{1}{n}\log P_n &\leq \frac{1}{n}\log\frac{2}{\sqrt{2\pi}} \\ &\quad -\frac{1}{n}\log\left|\frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{n}s_X}\right| \\ &\quad -\frac{1}{2n}\frac{\hat{\beta}_1^2}{\hat{\sigma}^2/ns_X^2} \end{aligned} \tag{52}$$

$$\begin{aligned} &= \frac{\log\sqrt{2}\pi}{n} \\ &\quad +\frac{\log\left|\frac{\hat{\beta}_1}{\hat{\sigma}/s_x}\right|}{n} \\ &\quad -\frac{\log n}{2n} \\ &\quad -\frac{\hat{\beta}_1^2}{2\hat{\sigma}^2/s_X^2} \end{aligned} \tag{53}$$

Take the limit as $n \to \infty$:

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\log P_n &\leq \lim_n \frac{\log\sqrt{2}\pi}{n} \\ &\quad +\lim_n \frac{\log\frac{\hat{\beta}_1}{\hat{\sigma}/s_x}}{n} \\ &\quad -\lim_n \frac{\log n}{2n} \\ &\quad -\lim_n \frac{\hat{\beta}_1^2}{2\hat{\sigma}^2/s_X^2} \end{aligned} \tag{54}$$

Since $\hat{\beta}_1/(\hat{\sigma}/s_X) \to \beta_1/(\sigma/s_X)$, and $n^{-1}\log n \to 0$,

$$\lim_{n\to\infty}\frac{1}{n}\log P_n \leq -\frac{\beta_1^2}{2\sigma^2/s_X^2} \tag{55}$$

I've only used the upper bound on $1-\Phi(x)$ from Eq. 48; if we use the lower bound from that equation, we get (Exercise 5)

$$\lim_{n\to\infty}\frac{1}{n}\log P_n \geq -\frac{\beta_1^2}{2\sigma^2/s_X^2} \tag{56}$$

Putting the upper and lower limits together,

$$\lim_{n\to\infty}\frac{1}{n}\log P_n = -\frac{\beta_1^2}{2\sigma^2/s_X^2}$$

Turn the limit around: at least for large n,

$$P_n \approx e^{-n\frac{\beta_1^2}{2\sigma^2/s_X^2}} \tag{57}$$

Thus, *any* $\beta_1 \neq 0$ will (eventually) give exponentially small p-values. This is why, as a saying among statisticians have it, "the p-value is a measure of sample size": any non-zero coefficient will become arbitrarily statistically significant with enough data. This is just another way of saying that with enough data, we can (and will) detect even arbitrarily small coefficients, which is what we *want*. The flip-side, however, is that it's just senseless to say that one coefficient is important because it has a really small p-value and another is unimportant because it's got a big p-value. As we can see from Eq. 57, the p-value runs together the magnitude of the coefficient ($|\beta_1|$), the sample size (n), the noise around the regression line (σ^2), and how spread out the data is along the x axis (s_X^2), the last of these because they control how precisely we can estimate β_1. Saying "this coefficient must be really important, because we can measure it really precisely" is not smart.

7 Confidence Sets and p-Values in R

When we estimate a model with `lm`, R makes it easy for us to extract the confidence intervals of the coefficients:

```
confint(object, level=0.95)
```

Here `object` is the name of the fitted model object, and `level` is the confidence level; if you want 95% confidence, you can omit that argument. For instance:

```
library(gamair); data(chicago)
death.temp.lm <- lm(death ~ tmpd, data=chicago)
confint(death.temp.lm)
```

```
##                   2.5 %     97.5 %
## (Intercept) 128.8783687 131.035734
## tmpd         -0.3096816  -0.269607
```

```
confint(death.temp.lm, level=0.90)
```

```
##                    5 %        95 %
## (Intercept) 129.0518426 130.8622598
## tmpd         -0.3064592  -0.2728294
```

If you want p-values for the coefficients[18], those are conveniently computed as part of the `summary` function:

[18] And, really, why do you?

```
coefficients(summary(death.temp.lm))
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 129.9570512 0.55022802 236.18763  0.00000e+00
## tmpd         -0.2896443 0.01022089 -28.33845 3.23449e-164
```

Notice how this actually gives us an array with four columns: the point estimate, the standard error, the t statistic, and finally the p-value. Each row corresponds to a different coefficient of the model. If we want, say, the p-value of the intercept, that's

```
coefficients(summary(death.temp.lm))[1,4]
```

```
## [1] 0
```

The summary function will also print out a *lot* of information about the model:

```
summary(death.temp.lm)
```

```
##
## Call:
## lm(formula = death ~ tmpd, data = chicago)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.275  -9.018  -0.754   8.187 305.952
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 129.95705    0.55023  236.19   <2e-16 ***
## tmpd         -0.28964    0.01022  -28.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.22 on 5112 degrees of freedom
## Multiple R-squared:  0.1358,Adjusted R-squared:  0.1356
## F-statistic: 803.1 on 1 and 5112 DF,  p-value: < 2.2e-16
```

As my use of `coefficients(summary(death.temp.lm))` above suggests, the `summary` function actually returns a complex object, which can be stored for later access, and printed. Controlling how it gets printed is done through the `print` function:

```
print(summary(death.temp.lm), signif.stars=FALSE, digits=3)
```

```
##
## Call:
## lm(formula = death ~ tmpd, data = chicago)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -42.27  -9.02  -0.75   8.19 305.95
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 129.9571     0.5502   236.2   <2e-16
## tmpd         -0.2896     0.0102   -28.3   <2e-16
##
## Residual standard error: 14.2 on 5112 degrees of freedom
## Multiple R-squared:  0.136,Adjusted R-squared:  0.136
## F-statistic:  803 on 1 and 5112 DF,  p-value: <2e-16
```

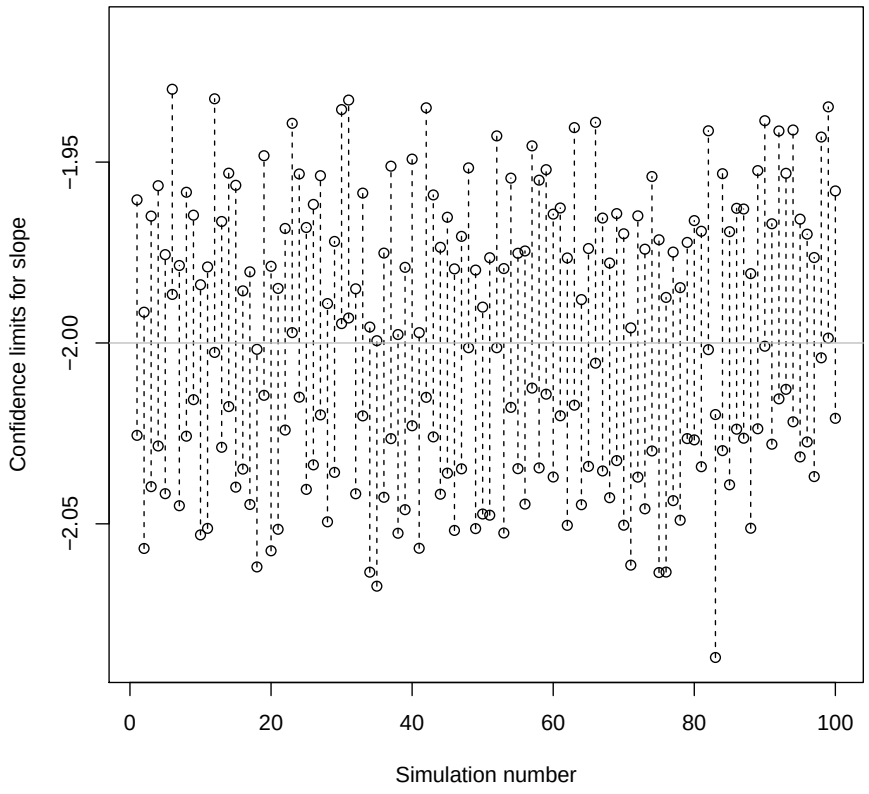
Here I am indulging in two of my pet peeves. It's been conventional (at least since the 1980s) to decorate this sort of regression output with stars beside the coefficients which are significant at various traditional levels. Since (as we've just seen at tedious length) statistical significance has almost nothing to do with real importance, this just clutters the print-out to no advantage[19]. Also, `summary` has a bad habit of using far more significant[20] digits than is justified by the precision of the estimates; I've reined that in.

7.1 Coverage of the Confidence Intervals: A Demo

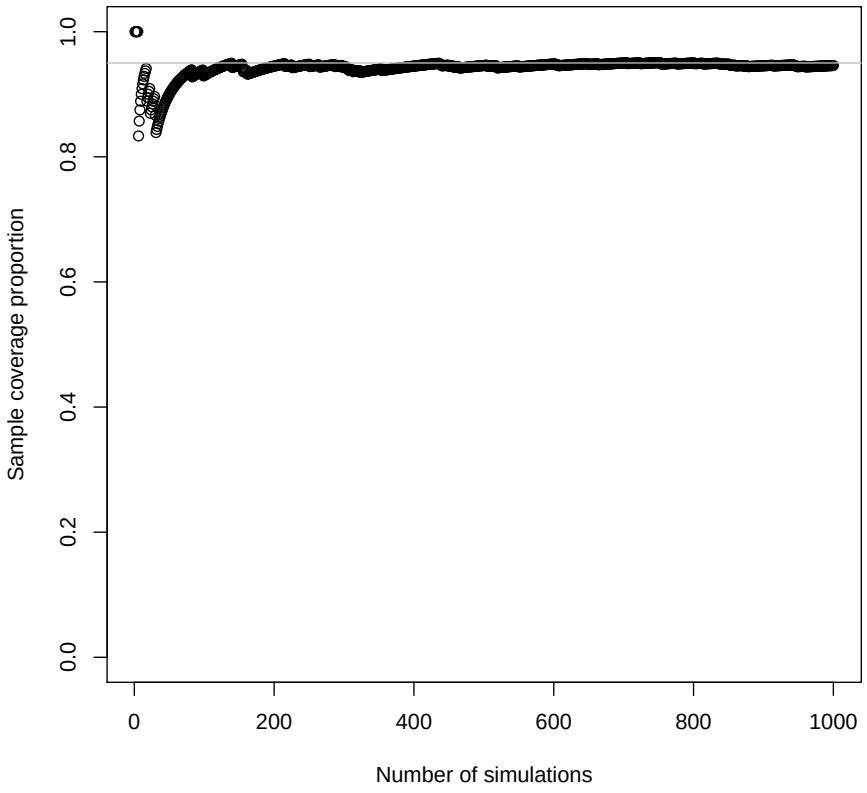
Here is a little computational demonstration of how the confidence interval for a parameter is a random parameter, and how it covers the true parameter value with the probability we want. I'll repeat many simulations of the model from Figure 2, calculate the confidence interval on each simulation, and plot those. I'll also keep track of how often, in the first m simulations, the confidence interval covers the truth; this should converge to $1 - \alpha$ as m grows.

[19] In fact, I strongly recommend running `options(show.signif.stars=FALSE)` at the beginning of your R script, to turn off the stars forever.

[20] A different sense of "significant"!




```
# Run 1000 simulations and get the confidence interval from each
CIs <- replicate(1000, confint(lm(y~x,data=sim.gnslrm(x=x,5,-2,0.1,FALSE)))[2,])
# Plot the first 100 confidence intervals; start with the lower limits
plot(1:100, CIs[1,1:100], ylim=c(min(CIs),max(CIs)),
     xlab="Simulation number", ylab="Confidence limits for slope")
# Now the lower limits
points(1:100, CIs[2,1:100])
# Draw line segments connecting them
segments(x0=1:100, x1=1:100, y0=CIs[1,1:100], y1=CIs[2,1:100], lty="dashed")
# Horizontal line at the true coefficient value
abline(h=-2, col="grey")
```




```
# For each simulation, check whether the interval covered the truth
covered <- (CIs[1,] <= -2) & (CIs[2,] >= -2)
# Calculate the cumulative proportion of simulations where the interval
# contained the truth, plot vs. number of simulations.
plot(1:length(covered), cumsum(covered)/(1:length(covered)),
     xlab="Number of simulations",
     ylab="Sample coverage proportion", ylim=c(0,1))
abline(h=0.95, col="grey")
```

8 Further Reading

There is a lot of literature on significance testing and p-values. They are often quite badly abused, leading to a harsh reaction against them, which in some cases goes as badly wrong as the abuses being complained of[21]. I find the work of D. Mayo and collaborators particularly useful here (Mayo, 1996; Mayo and Cox, 2006; Mayo and Spanos, 2006). You may also want to read `http://bactra.org/weblog/1111.html`, particularly if you find §6 interesting, or confusing.

The correspondence between confidence sets and hypothesis tests goes back to Neyman (1937), which was the first formal, conscious introduction of confidence sets. (As usual, there are precursors.) That every confidence set comes from inverting a hypothesis test is a classical result in statistical theory, which can be found in, e.g., Casella and Berger (2002). (See also Exercises 2 and 3 below.) Some confusion on this point seems to arise from people not realizing that "does $\hat{\beta}_1$ fall inside the sampling interval for β_1^*?" is a test of the hypothesis that $\beta_1 = \beta_1^*$.

In later lectures, we will look at how to get confidence sets for multiple parameters at once (when we do multiple linear regression), and how to get confidence sets by simulation, without assuming Gaussian noise (when we introduce the bootstrap).

Exercises

To think through or to practice on, not to hand in.

1. *Confidence interval for σ^2:* Start with the observation that $n\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-2}$.
   (a) Find a formula for the $1-\alpha$ sampling interval for $\hat{\sigma}^2$, in terms of the CDF of the χ^2_{n-2} distribution, α, n and σ^2. (Some of these might not appear in your answer.) Is the width of your sampling interval the same for all σ^2, the way the width of the sampling interval for $\hat{\beta}_1$ doesn't change with β_1?
   (b) Fix $\alpha = 0.05$, $n = 40$, and plot the sampling intervals against σ^2.
   (c) Find a formula for the $1-\alpha$ confidence interval for σ^2, in terms of $\hat{\sigma}^2$, the CDF of the χ^2_{n-2} distribution, α and n.
2. Suppose we start a way of testing the hypothesis $\beta = \beta^*$ which can be applied to any β^*, and which has size (false alarm / type I error) probability α for β^*. Show that the set of β retained by their tests is a confidence set, with confidence level $1-\alpha$. What happens if the size is $\leq \alpha$ for all β^* (rather than exactly α)?

[21] Look, for instance, at the exchange between McCloskey (2002); McCloskey and Ziliak (1996) and Hoover and Siegler (2008).

3. Suppose we start from a way of creating confidence sets which we know has confidence level $1-\alpha$. We test the hypothesis $\beta = \beta^*$ by rejecting when β^* is outside the confidence set, and retaining when β^* is inside the confidence set. Show that the size of this test is α. What happens if the initial confidence level is $\geq 1-\alpha$, rather exactly $1-\alpha$?

4. Prove that the p-value P is uniformly distributed under the null hypothesis. You may, throughout, assume that the test statistic T has a continuous distribution.

   (a) Show that if $Q \sim \text{Unif}(0,1)$, then $P = 1-Q$ has the same distribution.

   (b) Let X be a continuous random variable with CDF F. Show that $F(X) \sim \text{Unif}(0,1)$. *Hint:* the CDF of the uniform distribution $F_{\text{Unif}(0,1)}(x) = x$.

   (c) Show that P, as defined, is $1 - F_{|T|}(|T_{obs}|)$.

   (d) Using the previous parts, show that $P \sim \text{Unif}(0,1)$.

5. Use Eq. 48 to show Eq. 56, following the derivation of Eq. 55.

References

Casella, George and R. L. Berger (2002). *Statistical Inference*. Belmont, California: Duxbury Press, 2nd edn.

Feller, William (1957). *An Introduction to Probability Theory and Its Applications*, vol. I. New York: Wiley, 2nd edn.

Feynman, Richard P. (1985). *QED: The Strange Theory of Light and Matter*. Princeton, New Jersey: Princeton University Press.

Hoover, Kevin D. and Mark V. Siegler (2008). "Sound and Fury: McCloskey and Significance Testing in Economics." *Journal of Economic Methodology*, **15**: 1–37. URL `http://hdl.handle.net/10161/2045`. doi:10.1080/13501780801913298.

Mayo, Deborah G. (1996). *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.

Mayo, Deborah G. and D. R. Cox (2006). "Frequentist Statistics as a Theory of Inductive Inference." In *Optimality: The Second Erich L. Lehmann Symposium* (Javier Rojo, ed.), pp. 77–97. Bethesda, Maryland: Institute of Mathematical Statistics. URL `http://arxiv.org/abs/math.ST/0610846`.

Mayo, Deborah G. and Aris Spanos (2006). "Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction." *The British Journal for the Philosophy of Science*, **57**: 323–357. doi:10.1093/bjps/axl003.

McCloskey, D. N. (2002). *The Secret Sins of Economics*. Chicago: Prickly Paradigm Press. URL `www.prickly-paradigm.com/paradigm4.pdf`.

McCloskey, D. N. and S. T. Ziliak (1996). "The Standard Error of Regressions." *Journal of Economic Literature*, **34**: 97–114.

Neyman, Jerzy (1937). "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability." *Philosophical Transactions of the Royal Society A*, **236**: 333–380. doi:10.1098/rsta.1937.0005.