Lecture 4

One-sample problems

Disclaimer: *These lecture notes have not been subjected to the usual scrutiny reserved for formal publications. For instance, citations are mostly omitted or imprecisely made. The notes are meant as a memory aid for students of stat 305A at Stanford University. Copyright: A. B. Owen 2026.*

The prior lectures had philosophical and technical background on the linear model. Now we start looking at how to connect those ideas to applications. In practice, a statistical question usually begins with a discussion of domain ideas: yield of corn, whether somebody clicks, potency of a drug, return on a stock, associations with a gene and so on. It can take an hour's discussion, back and forth, to formulate their goals as a statistical problem.

The least squares story in this lecture is a linear regression with just an intercept. The applied statistics context here is to study hypothesis tests, p-values, and confidence intervals for a scalar. We look at how they are supposed to work and also look a bit at how they actually work. Many of those ideas go through to the more general single parameter settings such as those for a correlation coefficient.

4.1 Square one

The simplest setting for the linear model is where there aren't any x's to predict with. The linear model reduces to

$$y_i = \mu + \varepsilon_i. \tag{4.1}$$

If you like, $X \in \mathbb{R}^{n\times 1}$ is just a column of ones and $\beta \in \mathbb{R}^1$ got renamed as the scalar μ. The random variables ε_i are IID with mean zero and variance $\sigma^2 > 0$. We are mainly interested in estimating μ, but of course learning σ would also have value.

The least squares estimate of μ is obviously $\bar{y} = (1/n)\sum_{i=1}^n y_i$. The idea of averaging disparate numbers is second nature now but, according to Stigler (2016) it was once a radical idea. He describes contexts where a physical quantity was measured by multiple observers who all got different values. It seemed more natural then to many people to just reason about which observer was the most careful and use the one best value.

Where it was once controversial to view distinct observations as being sampled from a common random mechanism, now we often suppose that seemingly unrelated data sets are sampled from a random mechanism that makes whole data sets, one at a time. That leads to hierarchical models. In class, I used an example of the average age of abalone. It would not have been a radical idea to think that different abalone have different ages.

The intercept of interest is $\mu \in \mathbb{R}$. Much of what we say about the linear model for μ applies more generally to scalar parameters $\theta \in \mathbb{R}$ such as β_j to come later or correlation coefficients. So these notes merge analysis of an intercept with a general scalar parameter θ.

4.2 Staircase of inference

Suppose we measure 17 abalone ages, thinking that the population is possibly being over-harvested if the average age is below 10 years. These values come from the UCI repository:

$$15,\ 7,\ 9,\ 10,\ 7,\ 8,\ 20,\ 16,\ 9,\ 19,\ 14,\ 10,\ 11,\ 10,\ 10,\ 12,\ 7.$$

The sample average turns out to be 11.4. There is reasonable uncertainty about how that generalizes to the population we sampled from. We know that $\mathrm{var}(\bar{y}) = \sigma^2/n$ where σ^2 is the common variance of y_i. That very well describes the uncertainty in μ, but unfortunately we don't know σ^2. We can

however estimate σ^2 by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \quad \text{or} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

The first choice is a natural plug-in estimator and the second is the usual unbiased estimator. Now we might wonder whether s^2 is a good estimator of σ^2. We can estimate the variance of s^2 but it will depend on an unknown fourth moment quantity. We end up on Tukey's staircase of inference: each step is farther from the data using a harder to estimate 2^k-th moment of the distribution. E.g.,

$$\mathrm{var}(s^2) = \sigma^4\Big(\frac{2}{n-1} + \frac{\kappa}{n}\Big),$$

as can be found by grinding it out.

We could estimate κ using fourth sample moments, but that doesn't really settle things. The point of this example is that it is often harder to estimate the quality of an answer, like $\mathrm{var}(\hat{\mu})$ than to estimate the quantity we cared about in the first place, here μ. We almost always have to stop with some remaining uncertainty.

4.3 One sample t test

If $y_i \sim (\mu, \sigma^2)$ we might want to test whether $\mu = \mu_0$ is plausible for some specific value μ_0. Usually the value we want to test is 0. A zero parameter value in a more general regression means that some feature is perfectly unrelated to Y. In the abalone example, we might test $\mu_0 = 10$.

For this test we set up a null hypothesis $H_0 : \mu = \mu_0$. We will test this hypothesis either rejecting it or not rejecting it. Not rejecting it is called accepting but it does not mean that we believe H_0, much less have proved H_0. I suppose that acceptance might mean that H_0 is acceptable. At most.

There are exceptional cases where we somehow know that μ can only be one of two or a small list of values and there we might be able to establish one of those values by rejecting all the others. But usually we can only reject parameter values not establish them.

We juxtapose the null with an alternative hypothesis. It is commonly $H_A : \mu \neq \mu_0$.

In class I mentioned one-sided alternatives such as $H_A : \mu > \mu_0$. That is sometimes justified by reasoning that the only possible way for $\mu \neq \mu_0$ to arise is $\mu < \mu_0$. Maybe eating something is possibly harmful but could not

possibly be helpful. A second justification is that only $\mu < \mu_0$ is consequential and worth discovering. Maybe $\mu > \mu_0$ has no possible benefit or harm as for example an elevator cable that is stronger than specified.

This forces us to think a bit more about our domain. We want to get 'rewarded' for finding a discovery. We might do a one-sided test to favor the direction that we think is more likely to have a reward. Or it might favor the direction with the greater reward. It is about increasing the expected reward, with the first condition coming from a prior belief about where the reward might be.

It is possible to be wrong about both assertions and while that probability is not on the same footing as that driving random y_i, the chance of being wrong about these things may not be negligible compared to the statistical uncertainty we are working with. Also a common way that one-sided tests come up is post hoc. After failing to reject versus $\mu \neq \mu_0$ because $p = 0.06$, somebody might notice that one can reject versus $\mu > \mu_0$ at $p = 0.03$ and then reason afterwards that that was the only possible/consequential direction.

Using our distribution theory we know that, if $y_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ and if $\mu = \mu_0$, then

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \sim t_{(n-1)}. \tag{4.2}$$

Suppose for $n - 1 = 16$ degrees of freedom that we get the value 7.64 shown in Figure 4.1. That value looks to be incompatible with a t distribution. For the abalone data we might get $t = 1.438$, also marked on the figure. That one looks compatible.

From the t distribution $\Pr(t = 1.438)$ and $\Pr(t = 7.64)$ are both zero. The difference is that $\Pr(|t| \geqslant 7.64) \approx 10^{-6}$ while $\Pr(|t| \geqslant 1.438) \approx 0.17$. A proper test uses tail areas. The reason to use a tail area is to think of what would happen under H_A. Under H_A, $|t|$ would be much larger than it would be under H_0. Larger $|t|$ is stronger evidence against H_0. If we would have rejected for $|t| = 7.64$ then we would also reasonably reject for any $|t| > 7.64$.

For this problem we define a p-value as

$$p = \Pr\Big(|t_{(n-1)}| \geqslant \frac{|\bar{y} - \mu_0|}{s/\sqrt{n}}; H_0\Big).$$

The $t_{(n-1)}$ there is a random variable with the t distribution and the algebraic expression is the observed value (not the random variable that turned

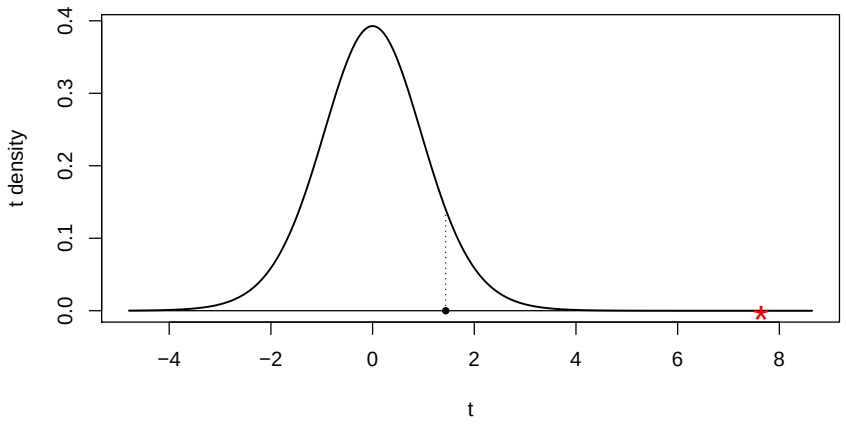


Figure 4.1: The curve is the $t_{(16)}$ density. The red mark is an unusually large value for this distribution. The solid point with a dotted line shows $t = 1.438$.

out to be the observed value). Ok, let's write it as

$$p = \Pr\Big(|t_{(n-1)}| \geqslant \frac{|\bar{y} - \mu_0|}{s/\sqrt{n}} \,|\, y_1, \ldots, y_n; H_0\Big),$$

to emphasize that $t_{(n-1)}$ is the only random quantity there. But we won't use that notation generally.

When p is small, like 10^{-6} we can reason that either H_0 is wrong or something of low probability happened. A super common error is to think that p is a probability that H_0 is false. We will say more about that later.

If H_0 is false then maybe $\mu \neq \mu_0$. Or maybe the data were not normal or not independent or not identically distributed.

For $t = 1.438$ we get $p = 0.17$. If $\mu = 10$, then there is about a 17% chance that $|t| \geqslant 1.438$. That casts no material doubt on H_0.

When p is small, say $p < \alpha$ for some threshold α, we might say that the result is statistically significant at level α. The original wording was about whether what we see in the data ***signifies*** anything at all about the world. If it could have just been 'random chance' then we should not draw strong conclusions from it.

A common threshold is $\alpha = 0.05$. That stems from Fisher's original paper. He contemplates other levels, but for his specific problem of considering whether agricultural yield is better after some innovation, he opted for $1/20$. In some genome wide association studies (GWAS) testing many gene loci for significance, the standard threshold is 5×10^{-8}.

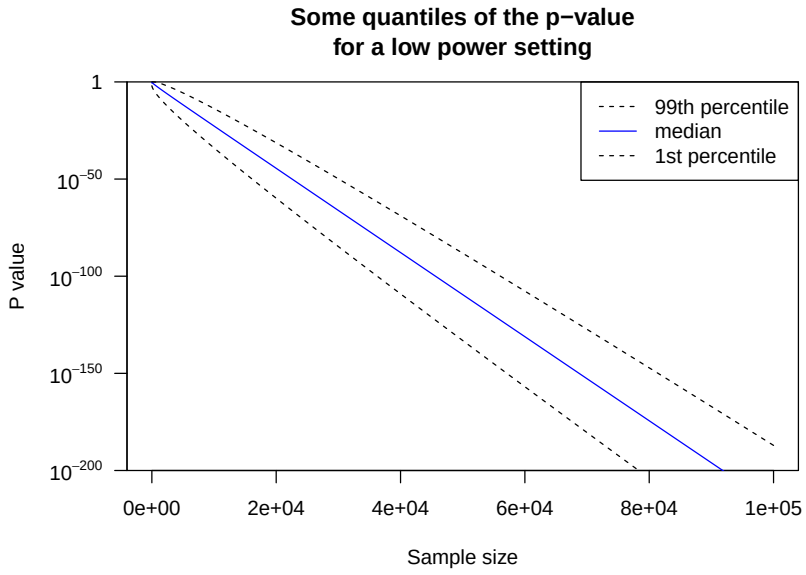


Figure 4.2: Here are quantiles of the p-value as a random variable when $|\mu - \mu_0|/\sigma = 0.1$, a small effect.

Handling a collection of p-values brings new issues. We will consider that in a later lecture.

4.4 Statistical vs practical significance

A tiny p-value like 10^{-88} can seem impressive. It might not be. It could arise from a small effect with an enormous sample size n. A hypothetical GWAS finding with $p < 5 \times 10^{-8}$ might correspond to a difference of 1mm of human height.

Richard Olshen has said that 'the p-value measures the sample size'. Even if $\mu_0 - \mu$ is quite small, so long as it is nonzero $p \to 0$ with n. Figure 4.2 shows how the p-value drops towards zero exponentially with n. This computation uses the noncentral F distribution that we saw in Lecture 2 and will see again in Section 4.6 where we discuss power.

Practical significance is about the magnitude of $|\mu - \mu_0|$. Our rough idea of it comes from $|\bar{y} - \mu_0|$. If $|\bar{y} - \mu_0|$ is large enough to matter for the application area then what we see is practically significant. In class we considered the 2×2 table of yes or no for practical and statistical significance,

		Practical Significance: Y	Practical Significance: N
Statistical Significance	Y	We have learned something important.	We got even more information than we needed.
	N	We should get more information. n was too small.	Even though we may not believe H_0 it may be accurate enough to use.

Table 4.1: Here are interpretations for a double dichotomy of statistical and practical significance.

in Table 4.1. Each of those four corners has a different consequence for us.

It can be tricky to select the value of δ wherein $|\mu_0 - \mu| \leqslant \delta$ is not practically significant and $|\mu_0 - \mu| > \delta$ is. This comes from the domain. It is pretty clear that $\mu_0 - \mu = 0$ is not practically significant, but beyond that, drawing the line requires considering downstream consequences of the use of the model. For example something that gives a driver 0.25 more miles per gallon might not make much difference to the driver but it might make a difference to the owner of a fleet.

Merely rejecting $H_0 : \theta = \theta_0$ is not necessarily a strong scientific finding. To quote Meehl (1990):

> In the social sciences and arguably in the biological sciences, "everything correlates to some extent with everything else."

Meehl (1997) argues in favor of confidence intervals over significance tests. Meehl had been making this point for many years prior to those two articles and he is certainly not alone in making this point. We won't have time to read everything written about it, but a statistical education should include at least some of these cautions.

p-values remain useful as a screening device. If we test 30,000 genes for association with some phenotype such as a health indicator, we might find that a much smaller number of them have statistically significant correlations. We may also only have the resources to do followups on a relatively small number of genes. It is then reasonable to ignore the non-significant ones, at least at first, and direct followup efforts to the significant ones.

4.5 Confidence intervals

Equation (4.2) is amazing. We have a formula with our desired unknown μ surrounded by quantities that we can compute. The combination is of course unknown but its distribution is known. It is called a ***pivotal*** quantity for this reason. Using the $1-\alpha/2$ quantile of the t distribution we get

$$1-\alpha = \Pr\Big(\frac{|\bar{y}-\mu|}{s/\sqrt{n}} \leqslant t_{(n-1)}^{1-\alpha/2}\Big).$$

We can manipulate this algebraically to get

$$1-\alpha = \Pr\Big(\bar{y} - t_{(n-1)}^{1-\alpha/2}\frac{s}{\sqrt{n}} \leqslant \mu \leqslant \bar{y} + t_{(n-1)}^{1-\alpha/2}\frac{s}{\sqrt{n}}\Big).$$

This is a $100(1-\alpha)\%$ confidence interval for μ.

This is exact. Tukey liked it for getting us out of the staircase of inference problem. But that's only if the data were normally distributed. We can of course test for normality or measure the amount of non-normality, but then that measure would have some kind of error. We're back on another staircase. Fortunately, the central limit theorem will be of use. The value $\bar{y}$ will be approximately normally distributed for large enough n, without requiring the y_i to be. Also, bootstrap confidence intervals coming later are effective.

A confidence interval takes the form

$$1-\alpha = \Pr(L \leqslant \mu \leqslant U). \tag{4.3}$$

In (4.3), it is the interval (L, U) that is random not the parameter μ. In a Bayesian analysis, coming later, this is reversed. There L and U are fixed (because we have observed them and conditioned on them) and the randomness models our knowledge of μ given the data.

Confidence intervals allow us to consider statistical and practical significance together. As noted above, the usual statistical context is about a scalar parameter θ, not necessarily $\mathbb{E}(y)$, and more usually a regression coefficient or a correlation. We can plot L, U, $\hat{\theta}$ and θ_0 on the same number line. If we have the idea that a practically significant difference from θ_0 would be δ units, then we can plot $\theta_0 \pm \delta$ on the same number line. If

$$(L, U) \subset [\theta_0 - \delta, \theta_0 + \delta]$$

then we can be confident that θ_0 is good enough to work with. Alternatively if

$$(L, U) \cap [\theta_0 - \delta, \theta_0 + \delta] = \varnothing$$

then we can be confident that θ differs from θ_0 by an important amount.

If the practical consequences of L and U are the same, then the remaining uncertainty about θ is not very important.

Figure 4.3 shows some ways that that a comparison of practical and statistical significance might work out. Each of the eight cases there could be the basis for a discussion, because the goals and costs can vary with the domain. If we have L and U each belonging to one of $(-\infty, \theta_0 - \delta)$, $(\theta_0 - \delta, \theta_0)$ $(\theta_0, \theta_0 + \delta)$, and $(\theta_0 + \delta, \infty)$ with $\hat{\theta} = (L + U)/2 > \theta_0$, then we need 6 different cases in Figure 4.3 to show them all. Two of the cases are duplicated to show 'close calls'.

We cannot enumerate all of the different discussions that might take place. The set of applied contexts is too varied. What we do see is that a simple dichotomy about whether to accept $\theta = \theta_0$ versus rejecting it in favor of $\theta = \hat{\theta}$ is too crude. Some examples follow.

In Fisher's time, θ might indicate the extra yield from changing an agricultural practice. There could be enormous costs in changing that practice and a delay until the next growing season to get new data. Then in order to make the change, one might reasonably want the whole confidence interval to be to the right of θ_0 to make it less likely that we have made a bad recommendation. Maybe the whole interval should even be beyond the level of practical significance, if that represents a switching cost.

In a modern A/B testing setting for electronic commerce, the cost to make a change could be modest. Also if $\theta < \theta_0$, then we may be able to find out quickly and make a correction. Then if we are wrong we would have small losses, and if we are right a potentially large gain. In that setting we might make a change if $\hat{\theta} > \theta_0$ with some statistical requirement less stringent than having the confidence interval exclude θ_0. It might just be about getting a positive expectation. Then rejecting H_0 is not necessary, though the decision may still consider the uncertainty in $\hat{\theta}$ via a confidence interval, or through Bayesian methods (that we consider later in the course).

Another possibility is that θ denotes the value of a physical constant, and θ_0 is the previously accepted value. We may be very cautious about taking an action like replacing θ_0 by $\hat{\theta}$ or similar.

The choice of a value δ denoting practical significance is not statistical, so it requires a discussion with a domain expert as alluded to above. I think that the necessity to do that is a large part of the reason that the literature emphasizes simply rejecting H_0 or not rejecting it. We can all agree that $\delta = 0$ is ***not*** an important difference but beyond that it gets difficult.

Confidence intervals and p-values have a dual property. Our t statistic based confidence set contains all of the values of θ_0 that would not have

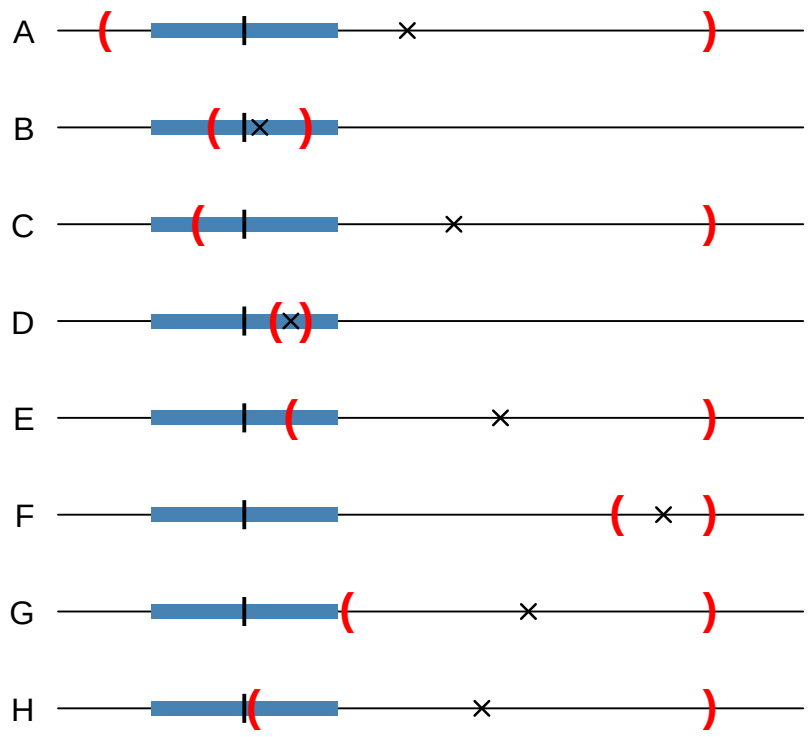


Figure 4.3: Here is a collection of possible confidence intervals for a problem where the null θ_0 (commonly zero) is marked as a black bar, and $\hat{\theta}$ is marked X. Somebody considers all values of θ within the horizontal blue bar centered at θ_0 to be practically equivalent to θ_0 and the large red parentheses mark confidence intervals. All of these cases have $\hat{\theta} > \theta_0$ and they all have $\hat{\theta}$ at the center of the confidence interval.

been rejected in a two sided test. Correspondingly, the points outside of the confidence set are the ones that are rejected. In more general settings a confidence set is typically an interval but does not have to be.

4.6 Power

We set up a test for $H_0 : \mu = \mu_0$ with an alternative $H_A : \mu \neq \mu_0$ and we reject H_0 when $|t| \geqslant t^{1-\alpha/2}_{(n-1)}$. We are naturally interested in the probability of this rejection. When $\mu \neq \mu_0$, the rejection probability is called ***power***. Using the relationship between t and F distributions, that follows directly

from their definitions, we get

$$\begin{aligned}\Pr(|t| > t^{1-\alpha/2}_{(n-1)}) &= \Pr(t^2 > F^{1-\alpha}_{1,n-1}) \\ &= \Pr\Big(\frac{n(\bar{y}-\mu_0)^2}{s^2} > F^{1-\alpha}_{1,n-1}\Big) \\ &= \Pr\left(\frac{\Big[\frac{\sqrt{n}(\bar{y}-\mu+\mu-\mu_0)}{\sigma}\Big]^2}{s^2/\sigma^2} > F^{1-\alpha}_{1,n-1}\right).\end{aligned}$$

Now

$$\frac{\sqrt{n}(\bar{y}-\mu+\mu-\mu_0)}{\sigma} \sim \mathcal{N}\Big(\sqrt{n}\frac{\mu-\mu_0}{\sigma}, 1\Big)$$

independently of $s^2/\sigma^2 \sim \chi^2_{(n-1)}/(n-1)$. Therefore t^2 has the noncentral F distribution $F'_{1,n-1}(\lambda)$ for

$$\lambda = \frac{n(\mu-\mu_0)^2}{\sigma^2} = \frac{(\mu-\mu_0)^2}{\sigma^2/n}.$$

The second expression writes the noncentrality parameter as a signal to noise ratio comparing squared signal to the variance of $\bar{y}$. We can also write $\lambda = n\Delta^2$ where $\Delta = (\mu-\mu_0)/\sigma$. The only way that μ, μ_0 and σ enter is through this ratio which is the number of standard deviations separating μ and μ_0. Only $|\Delta|$ enters when we use a two-sided alternative.

The power is then

$$1-\beta = \Pr\big(F'_{1,n-1}(\lambda) \geqslant F^{1-\alpha}_{1,n-1}\big), \tag{4.4}$$

where β denotes the probability of ***not*** rejecting H_0. Equation (4.4) relates α, β, μ, μ_0, σ and n. The most common usage is to select n to make β small enough.

Falsely rejecting H_0 when it actually holds is called a ***type I*** error, probably because it is the first thing we think of when creating a test. Failing to reject H_0 when it does not hold is called a ***type II*** error.

Let's do an example. Let $\mu_0 = 10$ and suppose that we think that a difference of size $|\mu-\mu_0| = 1$ is large enough to be important. We might test at $\alpha = 0.05$ and we would like $\beta = 0.05$ too, to get 95% power. In order to compute power for a given value of n, we need a value of σ. That is the hardest part, especially before we have gathered any data. It could be based on domain knowledge from similar problems but it might essentially be just a guess.

If we take $\sigma = 5$ and $n = 30$ we only get power 0.185, far short of our goal of 0.95. To get 0.95 we need $n \geqslant 327$. It is very common to compromise on power taking $\beta = 0.2$. For the given setting we would then need $n \geqslant 199$. This might still be expensive.

4.7 Non-normality

People have worried about the effects of non-normality for many years. Much of the work was done in the 1940s. Our t statistic satisfies

$$\frac{\sqrt{n}(\bar{y}-\mu)}{s} = \frac{\sqrt{n}(\bar{y}-\mu)/\sigma}{s/\sigma} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1).$$

If $\mathbb{E}(y^2) < \infty$ then the numerator converges in distribution to $\mathcal{N}(0,1)$ by the central limit theorem and the denominator converges to 1 by the strong law of large numbers. As a result the coverage error $\Pr(L \leqslant \mu \leqslant U) - (1-\alpha) \to 0$ as $n \to \infty$. We will see more about this when we consider the bootstrap.

If y_i have skewness γ and kurtosis κ then

$$\mathbb{E}(t) \approx -\frac{\gamma}{2\sqrt{n}}, \quad \mathrm{var}(t) \approx 1 + \frac{1}{n}\Big(2 + \frac{7}{4}\gamma^2\Big) \quad \text{and} \quad \gamma(t) \approx -\frac{2\gamma}{\sqrt{n}}.$$

These results are in Miller (1986). In this course we do not derive or even memorize these results. What we notice is that how non-Gaussian t is depends primarily on the skewness γ. That is also true for $\bar{y}$, but t has the additional complexity that both numerator and denominator are affected by outliers. It is notable that κ does ***not*** appear.

For an interesting exercise, work out what happens to t if we send $y_1 \to \infty$ (or $-\infty$) while holding $y_2, \ldots, y_n$ fixed. You can also sample y_i from the $t_{(2)}$ distribution which has finite mean but infinite variance and look at the coverage of the 95% Student's t based confidence interval for μ.

If $\gamma > 0$, then large positive observations dominate large negative ones. Roughly speaking: when $\bar{y}$ is larger than usual, so is s, but when $\bar{y}$ is smaller than usual s is not made larger. Then t dampens the large positive values of $\bar{y} - \mu$ more than it dampens the large negative ones.

4.8 Correlations

Correlations can be much worse for our t test than non-normality is. The damage they do is not ordinarily repaired by the central limit theorem and correlations are hard to detect statistically.

Suppose that $Y \sim \mathcal{N}(\mathbf{0}, R\sigma^2)$ where $\sigma > 0$ and $R \in \mathbb{R}^{n\times n}$ is a correlation matrix. That means it is a positive semi-definite matrix with diagonal entries $R_{ii} = 1$.

Here are a few simple models for R. We could have

$$R_{ij} = \begin{cases} 1, & i = j \\ \rho, & |i-j| = 1 \\ 0, & |i-j| > 1. \end{cases} \tag{4.5}$$

Then from our earlier notes we can work out that

$$\mathbb{E}(\bar{y}) = 0 \quad \text{and} \quad \mathrm{var}(\bar{y}) = \frac{\sigma^2}{n}\Big(1 + 2\rho\frac{n-1}{n}\Big)$$

while with a bit more work we can find that

$$\mathbb{E}(s^2) = \sigma^2\Big(1 - \frac{2\rho}{n}\Big).$$

For $\rho > 0$ and large n, the variance of $\hat{\mu}$ is about $1 + 2\rho$ times as large as we thought it was. At the same time our estimate of σ^2 is just a bit too small. We will get

$$t \xrightarrow{\mathrm{d}} \mathcal{N}(0, 1 + 2\rho).$$

The variance is higher than we think. We are naive by a factor of $1 + 2\rho$.

Puzzler: what about $\rho < -1/2$?

If the variance is more than we think, then our confidence interval is narrower than it should be. Similarly our p-values will be smaller than they should be. We would be over confident.

Another example has $R_{ij} = \rho^{|i-j|}$ for $-1 < \rho < 1$. Then the above argument shows that

$$t \xrightarrow{\mathrm{d}} \mathcal{N}\Big(0, \frac{1+\rho}{1-\rho}\Big).$$

Suppose that somehow our n data points got copy pasted into $2n$ observations. We think the variance is $\sigma^2/(2n)$. But really it is twice as large, σ^2/n. Our tests won't be reliable.

A more severe example has

$$R_{ij} = \begin{cases} 1, & i = j \\ \rho, & \text{else.} \end{cases}$$

You should work out how naive this is. For positive ρ it can be worse than copy-pasting your n observations to make $2n$ observations. This setting

is a model for election forecasting. The survey results in different states might have mean close to zero, but it is possible that they are all positively correlated. The total number of electoral college votes for one party will have much more variance than it would if the survey errors were uncorrelated.

4.9 Variance estimation

Sometimes we want to estimate σ^2. That is tricky when we have few degrees of freedom. One motivation for studying σ^2 is that for many products high quality is more about low variance than bias. I gave the example of table legs where variance might make a wobbly table while bias of the same magnitude might be unnoticeable.

Using our theory from an earlier lecture and assuming normality

$$\Pr\left(\frac{\chi^{2,\alpha/2}_{(n-1)}}{n-1} \leqslant \frac{s^2}{\sigma^2} \leqslant \frac{\chi^{2,1-\alpha/2}_{(n-1)}}{n-1}\right) = 1-\alpha,$$

where $\Pr(\chi^2_{(n-1)} \leqslant \chi^{2,p}_{(n-1)}) = p$, and we're using $p \in \{1-\alpha/2, \alpha/2\}$. It follows that

$$\Pr\left(\frac{s^2}{\chi^{2,1-\alpha/2}_{(n-1)}/(n-1)} \leqslant \sigma^2 \leqslant \frac{s^2}{\chi^{2,\alpha/2}_{(n-1)}/(n-1)}\right) = 1-\alpha,$$

gives us a $100(1-\alpha)\%$ confidence interval for σ^2. The upper limit uses small quantiles of $\chi^2_{(n-1)}$ because that captures the possibility that s^2/σ^2 was very small and needs to be multiplied by something larger than one.

When the degrees of freedom $n-1$ are very few then the upper confidence limit can be much larger than s^2. For $n = 3$ and 95% confidence it is about $39.5s^2$. For $n = 2$ we get $1018.25s^2$. Such tiny degrees of freedom are not commonly encountered but they can arise when data are very expensive to gather making n small.

4.10 Checking confidence intervals

We don't often get to check up on how confidence intervals do in real life. We can check predictions by how well they do matching y values that were not in our data set. For confidence intervals we rarely get the true parameter value.

Youden (1972) looked at confidence intervals for the astronomical unit (the distance from Earth to the sun). Figure 4.4 shows them versus the year

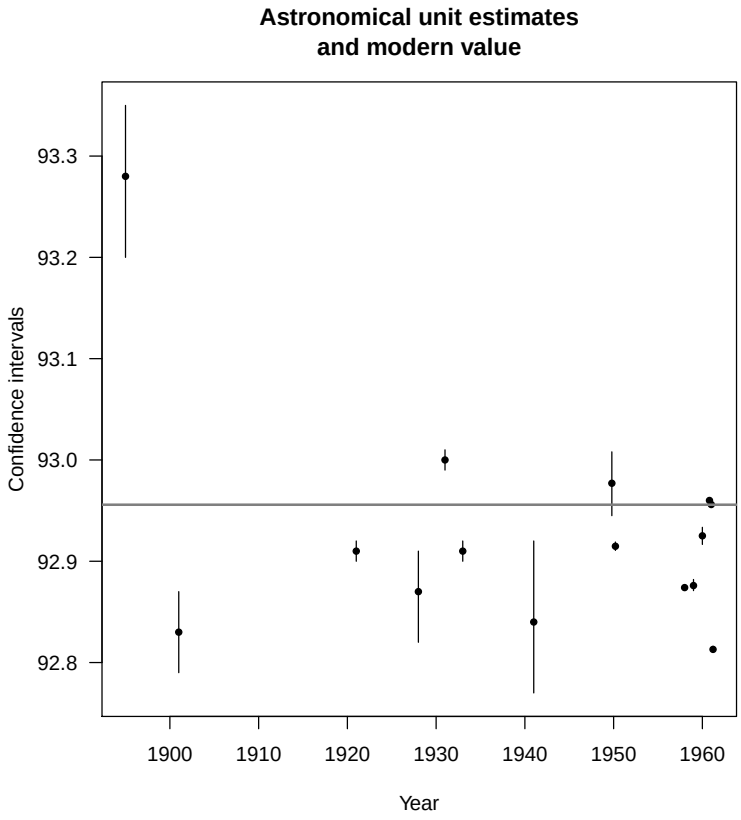


Figure 4.4: Successive confidence intervals for the astronomical unit (in millions of miles) collected by Youden.

they came out comparing them to the modern value. No interval contained the following center point. Maybe one or two covered the modern value. The confidence intervals quantified the uncertainty that the users knew about. There may have been some remaining bias in their estimates or other sources of uncertainty.

4.11 Low power

We are very often interested in whether some scalar parameter θ is not zero. In the prior sections that was $\theta = \mu - \mu_0$, but more often θ is something like a regression coefficient β_j or a correlation coefficient and the value zero is

very special because it means that some predictor does not affect the mean of y and maybe doesn't affect y in any way at all.

While we often believe that θ could be small, there's almost no possibility that it is exactly zero. So it is then a puzzle why we should test $H_0 : \theta = 0$ when we're sure $\theta \neq 0$. George Barnard reasoned that if we cannot reject $H_0 : \theta = 0$ then we cannot know much about $\text{sign}(\theta)$.

There was an interesting blog post by Andrew Gelman in 2014 that showed that in a setting with low power, even if we do reject H_0, we could still have $\text{sign}(\hat{\theta}) \neq \text{sign}(\theta)$. He wrote about what happens with power 6%. His numerical example was more like 5.5%, but the larger point still holds, despite that error.

Why would we have low power? In a company running A/B tests to improve their operations, tiny effects could be worth a very large multiple of the salaries paid to the people trying to improve things. They might be close to a local optimum. An industrial product might be very expensive and the testing could be destructive, so n is small. Sometimes our data have many positive off-diagonal correlations and those reduce power. We look at this in Section 4.8.

We can study this situation by supposing that the effect $\theta = 1$ and we test at $\alpha = 0.05$. To have power 0.06 we could have had $\sigma/\sqrt{n} = 3.39$ (in a setting where n was large enough that the t distribution is nearly Gaussian and then σ was large). Then rejection happens when and only when $|\hat{\theta}| \geqslant 1.96 \times 3.39 = 6.65$. The true effect is 1. Any rejection whatsoever exaggerates it by more than 6-fold. Figure 4.5 illustrates the situation.

Any time we reject H_0 in a low power setting, we get a greatly exaggerated effect size. Figure 4.6 shows 100 confidence intervals that exclude 0 when $\theta = 1$, $\alpha = 0.05$ and the power is 0.06. They all are centered on values $|\hat{\theta}| \geqslant |\theta| = 1$, exaggerating the effect size. We should not expect that to replicate if we run another study. This effect is known as the ***winner's curse***.

4.12 Probability that $\theta \neq 0$

A common misconception is that $p = \Pr(H_0 \text{ is false} \,|\, y_1, \ldots, y_n)$ while it is really just the probability that $|\hat{\theta} - \theta_0|$ would be as large as or larger than what we saw, if H_0 were true. I would actually rather have $\Pr(H_0 \text{ is false} \,|\, y_1, \ldots, y_n)$ than a p-value.

We can get a conclusion like that from a Bayesian analysis. We will look at Bayesian methods in a later lecture. The cost is that you have to

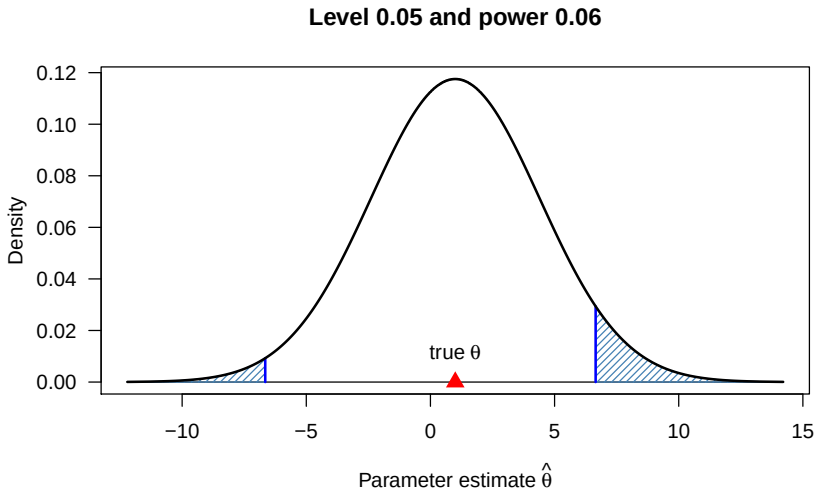


Figure 4.5: This figure illustrates a setting where power is 0.06 for a two-sided test at $\alpha = 0.05$. The triangle marks the true effect size $\theta = 1$. There is a $\mathcal{N}(\theta, 3.39^2)$ density for $\hat{\theta}$. The shaded areas illustrate rejections. About 20% of the rejections have the wrong sign.

specify a model in which θ is drawn from some distribution and then we observe Y given θ from some conditional distribution. With those in place we can compute the posterior distribution of θ given Y. The conclusion is better but the cost is higher; we take on an additional task with its own uncertainty.

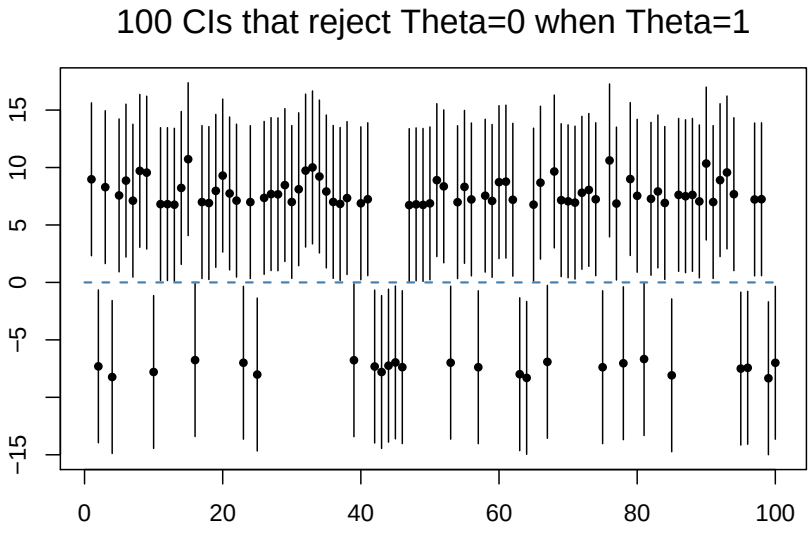


Figure 4.6: This figure shows 100 low power confidence intervals that rejected H_0 when $\theta = 1$. All the effect sizes are greatly exaggerated. About 20% of the signs are wrong. The edges of the confidence intervals are close to zero.

Bibliography

P. E. Meehl. Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, 66:195–244, 1990.

P. E. Meehl. The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik, and J. H. Steiger, editors, *What If There Were No Significance Tests?*, pages 393–425. Lawrence Erlbaum Associates, Mahwah, NJ, 1997.

R. G. Miller. *Beyond ANOVA, Basics of Applied Statistics*. Wiley, 1986.

S. M. Stigler. *The seven pillars of statistical wisdom*. Harvard University Press, 2016.

W. J. Youden. Enduring values. *Technometrics*, 14(1):1–11, 1972.