

1 Overview

2 Confidence Intervals

3 Chi-square and t- Distributions

4 Confidence Intervals: Unknown Population Variance

Week 4 Notes Part 3 Confidence Intervals

Code ▾

Author: Brendan Gongol

Last update: 07 January, 2023

γ = confidence level.

95% confidence interval == $\gamma = 0.95$

1 Overview

In this part of the notes, we will introduce confidence intervals. First we will illustrate the concept in a simpler case of known underlying variance; then we will introduce Student t distribution and describe how confidence intervals are computed in the more realistic case when the variance estimate is also determined from the sample.

All of section 2 assumes we know the variance ^{population} ~~var~~

2 Confidence Intervals

Instead of just saying significant or not significant CI's give you a range of values, where the value of interest (eg. true mean) would lie between at a confidence (say 95%).

So far we have been using hypothesis testing framework in order to quantify the significance of our findings. Within that framework, upon choosing some significance cutoff, be that 5% or 0.1%, we make a single "yes/no" call: the finding observed in a given sample is either significant or not.

While hypothesis testing provides an extremely useful and universal framework, we can also ask a different question:

- Given the observations from a given sample, what is the range, which likely contains the true underlying value of some sample statistics?

Consider the mean as an example. We know that the sample mean is a random variable and changes from sample to sample. We also know that it is expected to fluctuate around the true underlying mean. Suppose in a given sample we observed a sample mean $\mu=0.4$. We can be pretty sure that this is not exactly the true underlying mean, but we can hope that it is "close enough". Can we use the data in order to actually define an interval $[\mu_{low}, \mu_{high}]$ such that the true underlying mean would be found in this interval with "probability" equal to, e.g. 95%? Such interval is what we call a 95% confidence interval. [Beware of the above use of the word "probability"!! We used somewhat colloquial language here, it might be more intuitive but remember that the underlying mean (at least in frequentist interpretation of statistics) is not a random variable. It is fixed (although it might be unknown to us) so there is no such thing as "probability" for that value to be anywhere. We will get to a more accurate definition later]

the idea is to create a range $[\bar{x}_{low}, \bar{x}_{high}]$ such that we can be "confident" (say, 95% confident) that the true mean (μ) lies within that range.

Let us work through our example a little more and get into some detail. In Part 2 of this week's notes we understood that for the sample of size n drawn from the population with mean μ and (known!) variance σ , the sample mean is distributed as

1 The CLT tells us that the sample mean (\bar{x}) is approx normally distributed.

known population variance.

* remember the SD of

\bar{x} is $\frac{\sigma}{\sqrt{n}} = \text{SEM}$.

$$f(\bar{x}) \propto \exp\left(\frac{-(\bar{x} - \mu)^2}{2(\sigma/\sqrt{n})^2}\right)$$

where we omitted the normalization factor in front of the exponent (it's not of interest for us right now), and used the fact that the standard deviation of the sample mean is σ/\sqrt{n} . Let us introduce new variable,

2

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

purpose of z score = converts \bar{x} into a standardized form whose distribution doesn't rely on μ , σ , or n .

This variable (often called Z-score in the literature) measures the deviation of a random variable from its mean in units of the standard deviation of that variable. For instance, if $Z = 1$, then variable \bar{x} is one standard deviation from its expected value μ (one standard deviation of variable \bar{x} itself, of course, which is σ/\sqrt{n} , not standard deviation σ of the underlying variable X). Substituting definition (2) into the distribution (1), we can immediately see that if \bar{x} follows normal distribution (1), then Z follows a standardized normal distribution (i.e. normal with $\mu = 0$ and $\sigma = 1$): $N(0, 1)$.

3

$$f(Z) = \frac{1}{\sqrt{2\pi}} \exp(-Z^2/2)$$

This is nothing but a simple variable substitution. But what was achieved by it is that now we have the universal distribution that does not depend on μ , σ or n anymore. For any gaussian variable, if we express its values in terms of deviations from the mean, scaled by the sd (i.e. use Z-score), the distributions of all such variables (in Z-score terms) will be all the same - the distribution (3) above. Hence we can setup some universal thresholds based on (3). For instance, for normal distribution, 95% of its mass is concentrated within 1.96 standard deviations on each side of the mean, so in case of the standardized distribution (3), with $\mu = 0$ and $\sigma = 1$, we simply have

4

$$P(-z_o < Z < z_o) = \gamma = 0.95 \text{ if } z_o = 1.96$$

where $P(a < Z < b)$ is the probability to find Z within the interval $[a, b]$, i.e. $\int_a^b f(Z) dZ$. Now let us consider again expression (2). Substituting it back into (4), we obtain that for any distribution with given μ and σ and given the sufficiently large sample size n (such that CLT works, the mean is nearly normally distributed and hence (1) and (4) apply), the sample mean distribution is such that

5


← this is the 95% CI for the true mean μ .

$$P(-1.96 < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < 1.96) = 0.95$$

This probability should be interpreted as follows: in 95% of cases (i.e. samples), the rescaled deviation of \bar{x} from μ is within the specified boundaries. Solving for μ we obtain the 95% confidence interval:

$$\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}$$

A few important notes are due:

- The constant 1.96 came from standardized distribution (3) and the requirement of 95% of its mass being concentrated within $[-z_0, z_0]$ interval (Eq.(4)). If we were looking for, e.g., 79% or 98% confidence intervals, we would have to find the corresponding z_0 such that 0.79 or 0.98, respectively, of the standardized normal distribution's mass is concentrated within $[-z_0, z_0]$, similar to (4). So we would have some other value for z_0 in those cases. Importantly, those values would be just numbers (because (3) is standardized and does not contain any parameters!). As a result, we would simply have a different constant instead of 1.96 in the confidence interval (6).
- The confidence interval can be (casually) interpreted as follows: given the measurement (a sample), it is a range, where we expect the true underlying mean to be located (with some limited confidence, be that 90 or 99% - we can define a confidence interval with the confidence level that suits our need as described in the previous paragraph).
-  Note that the confidence interval is an estimate (random variable) in itself: it depends on the sample mean \bar{x} . So if we draw next sample, it will have a different \bar{x} and hence different confidence interval.
- This brings us to the accurate way of interpreting confidence intervals. In "classical", "frequentist" statistics, if we have a sample, we compute its sample mean \bar{x} and then compute the confidence interval (6). While we keep re-iterating that these are random variables themselves, this refers to future measurements: if we keep drawing samples, we will observe fluctuations from sample to sample. But both \bar{x} and the confidence interval obtained from a given sample we are looking at are fixed values (they are already "realized"), there is no probability associated with them, as these are certain events that happened in the past. The confidence interval (6) either contains or does not contain the true underlying mean, again without any associated probability, this is a certain event (we just do not know which possibility has been realized). What (6) tells us is that if we keep drawing multiple samples and calculate the sample mean and the confidence interval (6) for each of them, then 95% of such intervals (samples) will encompass the true underlying mean.
- The above interpretation of the confidence interval should not be too surprising, if you think about it. Consider the more familiar case of hypothesis testing. We obtain sample average \bar{x} for a given sample. This measurement happened, so it's not probabilistic, it's a fact of life. This sample mean can be close to the underlying population mean or we could have got an unlucky sample draw so that the sample mean is not very representative. Despite the fact that one or the other possibility is already realized with certainty, we do not know which one it was. The only argument we could put forward is the one involving multiple sampling: e.g., if the underlying mean is zero, and if we draw multiple samples, how many of them would give us sample mean at least as large as \bar{x} (cf.: if we draw multiple samples and compute confidence interval for each of them, how many of these confidence intervals contain the true underlying mean?). If many such repeated samples would indeed give \bar{x} or larger for their sample mean, we must stay with null hypothesis ($\mu = 0$), since we have no evidence to the contrary (cf.: our confidence interval contains 0 and if we draw multiple samples, 95% of the intervals computed according to exactly the same prescription are expected to include the true mean, so we must stay with $\mu=0$ assumption as well, as we got no compelling evidence against it)

As a matter of fact, as you could guess by now, confidence intervals and hypothesis testing are very closely related. Depending on the way the estimations are set up, they can be completely equivalent. For instance, let's take another look at our simple example. Suppose we are interested in the usual test for the underlying population mean equal to zero. If we were to use hypothesis testing, we would ask about the probability to observe sample mean equal to or greater than x (by absolute value in two-sided test) under the condition $\mu=0$. If we use a standardized representation (2), then we have (assuming $\mu = 0$)

7

$$\bar{z}_0 = \frac{\bar{x}}{\sigma/\sqrt{n}}$$

and the probability to observe (across multiple samplings) a more extreme value is $P(|Z| > \bar{z}_0)$ (assuming for the sake of certainty that $\bar{z}_0 > 0$), it's the same probability function derived from standardized normal as in (4). If we set a significance cutoff α , then we reject the null (i.e. call our observed sample mean "significant") when

8 ~~the probability that the deviation of a random var from its mean is greater~~

Remember Z = deviation of a random variable from its mean.

* see written notes pt 8.

$$P(|Z| > \bar{z}_0) < \alpha$$

The probability in the left side of last expression ("probability to observe any value Z above \bar{z}_0 , by abs. value") is a function of only the observed (normalized) sample mean \bar{z}_0 . It is easy to see that this probability is the area under the tails of the normal distribution (3), starting from \bar{z}_0 , $2 \int_{\bar{z}_0}^{\infty} \exp(-Z^2/2) dZ / \sqrt{2\pi}$ (the factor of 2 accounts for positive and negative tails), so it is a monotonously decreasing function: when \bar{z}_0 increases, the probability to observe any value above it obviously decreases. Hence, if we find a value \bar{z}_0 such that

9

$$P(|Z| > z_0) = \alpha$$

Then for all values $\bar{z}_0 > z_0$ the condition (8) will hold, in other words any \bar{z}_0 above the threshold z_0 would be significant in our hypothesis testing procedure. Remember that we are still looking at the case here, where we know the exact value of the standard deviation. So it's not too surprising that there is a cutoff on the x itself: when we are at or above specific distance from the mean (in terms of known standard deviation), the probability to "be there" also drops below some fixed threshold. Using (7), we can rephrase this statement in terms of the sample mean itself: at any given confidence threshold α , any observed sample mean \bar{x} such that

10

$$\bar{x} > z_0 \frac{\sigma}{\sqrt{n}}$$

will be significant, where z_0 is determined from solution of Eq. (9).

Let's now look closely at Eq. (4). It is also an equation for $0 z$ (above we cited a solution for $\gamma = 0.95$ (which gives $z_0 = 1.96$), but we can solve it (at least in principle) for any γ . Note that

11

the total probability of all outcomes must equal 1.

$$P(-z_0 < Z < z_0) + P(|Z| > z_0) = 1$$

probability that R.Variable (z-score) falls w/in the range

$[-z_0, z_0]$ - It represents the middle of the

the $|Z| > z_0$ notation captures both directions (absolute value)
probability that R.V Z falls outside the range $[-z_0, z_0]$, either in the left tail ($Z < -z_0$) or right tail ($Z > z_0$).

(the probability to observe any value of Z is one!). Hence, solving (9) for a fixed value α or solving (4) for $\gamma = 1 - \alpha$ will result in exactly the same value of z_o . Using this value of z_o in (8) instead of 1.96 (which was specific solution for $\gamma=0.95$), we get confidence interval

12

$$\bar{x} - z_o \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_o \frac{\sigma}{\sqrt{n}}$$

Let us now compare (12) and (10): when our sample mean satisfies (10) (i.e. we consider the finding significant), the left boundary of the interval (12) computed for the complementary level of significance $\gamma = 1 - \alpha$ (i.e. the same z_o) is positive, i.e. the confidence interval does not contain $\mu = 0$ (which value was the specific assumption of our null hypothesis). In contrast, for sample means that are not significant according to any given significance threshold α , the corresponding confidence interval at the level of significance $\gamma = 1 - \alpha$ will span over $\mu = 0$ (so in a sense, confidence interval will be telling us that the null hypothesis value $\mu = 0$ is still plausible). Hence in this case the hypothesis testing and (exact) confidence interval convey the same information regarding the acceptance/rejection of the null. Note that while hypothesis testing gives a binary yes/no (accept/reject) answer, the confidence interval informs us of the whole range where underlying mean could lie and still (most likely) result in the observed sample mean.

3 Chi-square and t- Distributions

In the previous section we have developed all the required machinery, but we were specifically using the normal distribution (or the sample mean) in all our examples. All the equations and conditions in the previous section that involve probabilities P to observe any value above or below some threshold value z_o are generic for any variable Z (i.e. for any distribution $f(Z)$); but the specific values (e.g. $z_o = 1.96$ as a solution of $P(-z_o < Z < z_o) = 0.95$) are of course unique to the normal distribution (3). In general,

$$P(-z_o < Z < z_o) = \int_{-z_o}^{z_o} f(Z) dZ$$

for any distribution $f(Z)$, and this equation should be solved for z_o (most likely numerically, in most cases analytical solution will not be possible).

The CLT does tell us that the distribution of the sample mean is normal (or at least approaches normal with increasing sample size n), but in order to use that distribution we need to know the standard deviation of the underlying population σ (as used in all the equations in the previous section, e.g. in the formula for the confidence interval (6) or (12)). In reality we usually do not know σ and instead estimate it from the same sample, i.e. we use the sample standard deviation s as an approximation. In principle, we could simply substitute s for σ in all the final results (i.e. we still use normal distribution (3) in order to determine z_o from condition (4), and then simply use s in the final expression (12)). This would indeed give us some reasonable approximation. However there is a more accurate (and principled) way of solving this problem.

Namely, instead of (2), let us consider random variable

13

s = sample SD

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

$$T = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Remember the types of SD:

σ = population SD = spread of data in whole population

s = sample SD = measures spread within a sample taken from the population

\underline{s} is an estimator of σ

$\sigma_{\bar{x}}$ = SEM = standard error of the mean measures variability of the \bar{x} across all samples drawn from P.P.

i.e. we use the estimate of the standard deviation from the start on and measure deviation of the sample mean from the (unknown) population mean in units of s/\sqrt{n} . Now we have a fraction of two random variables (since both \bar{x} and s are random variables). In the numerator, we have \bar{x} , which is still normally distributed, of course, at least for large n (CLT still holds). Note that if we do assume right away that the underlying distribution is normal (as t-test actually does), then \bar{x} is normally distributed at any value of n , even small ones (because the sum of any number of normal variables is normal). But Eq.(13) also contains another random variable, s , in the denominator. When we draw next sample, the values of both \bar{x} and s change, and the probabilities to observe any particular value of these two variables are governed by their own distributions. What is the distribution of the fraction? *Because both \bar{x} and s are random, T does not follow a normal distribution, instead it follows Student's ~~normal~~ t-distribution.*

Detailed characterization of these distributions is beyond the scope of this course, however these distributions are extremely important and you have to be aware of them and know where they come from.

First, consider a random variable that is a sum of squares of independent variables X_i : *① the equation to find s^2 depends on the sum of squared deviations of x from \bar{x} .*

14

$$Y = X_1^2 + \dots + X_n^2$$

② the SoS of independent Variables X_i - the sample variance (s^2) heavily relies on the distribution of X_i . \rightarrow when X is normally distributed

With normally distributed X , random variable Y follows the distribution, known as chi-square:

$$Y \sim \chi^2(n) \text{ or } Y \sim \chi_n^2$$

(shorthand: $X \sim N(\mu, \sigma)$),

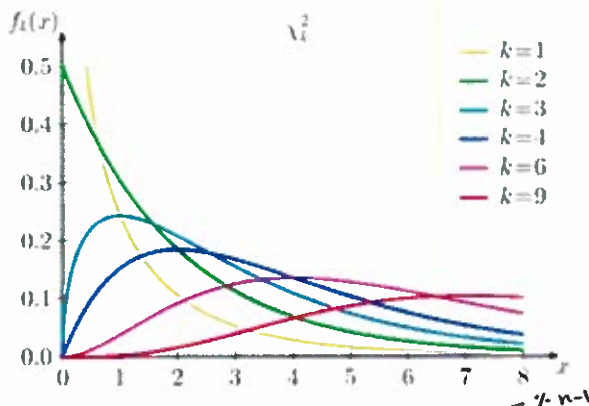
with the probability density function

the sum of squares follows a chi-squared distribution. B/c of this s^2 is also a random variable.

15

$$f_n(x) = \begin{cases} \frac{x^{n/2-1} e^{-x/2}}{2^{n/2} \Gamma(n/2)} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Note that since Y is a sum of i.i.d. random variables, according to CLT at large n the distribution of Y will be approaching normal regardless of the distribution function of the summands. In other words, chi-square distribution (15) must converge to normal as n increases. The plots of chi-square distributions for few different values of n are shown below (credit: Wikipedia); note how at small n the distribution looks very differently from a normal one, but with increasing n it moves away from zero and looks more and more gaussian-like.



A normal variable \div by a chi-square dependent variable leads to T following the Student's t-distribution

When n is small: the variability of s is significant, leading to heavier tails in the t-distribution.

When n is large: The chi-square distribution converges toward a normal distribution (by the CLT) and T approaches the standard normal distribution ($N(0,1)$).

Since the sample variance is the (scaled) sum of squares of i.i.d. random variables, the standard normal

$$s^2 = (x_1^2 + \dots + x_n^2) / (n - 1)$$

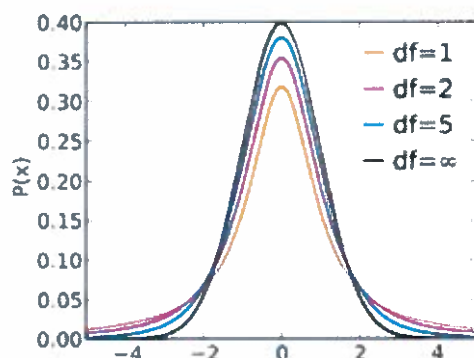
(assuming $\mu = 0$), it is thus distributed as chi-square if the random variable X is normally distributed.

It can be further shown that the random variable T defined as a ratio of normally distributed variable to a chi-square variable, as in (13), follows so-called Student's t-distribution:

16

$$f_n(t) = \frac{1}{\sqrt{n}B(1/2, n/2)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

where B is the beta-function. Important: for the sample of size n , its t-statistics (13) follows t-distribution (16) with $n-1$ degrees of freedom. Examples of t-distributions at different n (referred to as "degrees of freedom") are shown below (source: Wikipedia).



Student's distribution is exactly what is used by t-test: the value of statistics T for a given sample is defined using (13), and then exactly the same procedure is followed as the one we have demonstrated in Part 2 (Eq. 6) or in the previous section, but using correct distribution of T (t-distribution shown above, 16), instead of normal.

We can also see now where the assumptions and requirements of the t-test come from: t-distribution describes the probability density of a ratio of a normal and chi-square variables.

- For small n , the sample mean (numerator in (13)) is normal only when X is normal (CLT will not kick in at very small n , it promises only asymptotic behavior)
- The sample variance has chi-square distribution only when X is normal;
- At large sample sizes, the sample mean is normally distributed due to CLT, and so is sample variance (14), so the t-distribution (and t-test) can be expected to work reasonably well, even when X is non-normal.

4 Confidence Intervals: Unknown Population Variance

This section don't know population μ or σ^2 .

Now we are finally in position to build the confidence interval in a more realistic case when we have a sample but we do not know the underlying population variance, so we have to estimate it from the sample itself.

Putting together all the machinery we just developed: first we define dimensionless variable T as in (13). It should follow a t-distribution (if variable X is normal or n is sufficiently large!). Hence we should use the t-distribution when solving (4). In R, many functions will do it for us automatically, but if we want to compute the confidence interval manually (and you will be asked to in your homework!), we can make the following observation.

If X is normal or n is large then T follows the t-distribution. So when solving for CIs we use the T distribution instead of normal distribution.

When looking for confidence interval at significance level γ , we need, similarly to Eq. (4), to find the boundary z_o such that for the t-distribution the probability to observe the value T in the interval $[-z_o, z_o]$ is equal to γ .

17

$$P(-z_o < T < z_o) = \gamma$$

The t-distribution is symmetric, so out of the total probability $1 - \gamma$ to find T outside of the range, half of the cases will be on the left

18

$$P(T < -z_o) = (1 - \gamma)/2$$

Step 1: find boundaries according to our CI value of interest ($\gamma = 0.95$)

and half will be on the right

19

$$P(T < z_o) = (1 - \gamma)/2$$

Or equivalently,

20

$$P(T < z_o) = 1 - P(T > z_o) = 1 - (1 - \gamma)/2$$

Solving (17) for z_o is equivalent to solving (18)-(20), all these equations are just equivalent re-formulations of (17). But what is the value $-z_o$ such that for a given distribution the probability to observe a value below $-z_o$ (Eq.18) is equal to the $(1 - \gamma)/2$? This is by definition the lowest $(1 - \gamma)/2$ -th quantile of that distribution. Similarly, (20) can be interpreted as the highest $1 - (1 - \gamma)/2$ -th quantile of the distribution. z_o For instance, if we are looking for the 95% confidence interval ($\gamma=0.95$), then we want the boundary z_o such that 95% of the distribution is within $[-z_o, z_o]$, or equivalently, we want to leave out 2.5% of the distribution on the low and higher ends, so the interval $[-z_o, z_o]$ is actually [2.5% quantile, 97.5% quantile]. This is something we can easily simulate and compute in R. After the boundaries z_o of the interval for T are determined according to the specified significance level γ , we can use (13) in order to calculate actual boundaries of the confidence interval:

$$-z_o < T = \frac{\bar{x} - \mu}{s/\sqrt{n}} < z_o$$

Hence,

21

$$\bar{x} - z_o \frac{s}{\sqrt{n}} < \mu < \bar{x} + z_o \frac{s}{\sqrt{n}}$$

Step 2: rearrange formula to solve for pop mean (μ) with our boundaries $(-z_o, z_o)$ that we have calculated.

Note that while (21) looks just the same as (12), only with σ replaced with s , these two expressions actually use different values of z_o . In Section 2 we assumed normal distribution of Z (σ was known), so that z_o was determined by the quantiles of the normal distribution; here we used quantiles of a t-distribution in order to obtain z_o .

Let us illustrate the material above with a simple calculation in R. First we are going to randomly draw a sample from a normal distribution and compute the confidence interval of the mean. In order to do that, we will use a little trick: there is a function `confint()` in R that computes confidence intervals of the coefficients in an object representing a model fit. It would be convenient to use this function, but we need to build the fitted model object first. However, if we calculate the best “linear” fit using the model $X \sim 1$, i.e. intercept only, slope=0, the intercept of such a model is obviously going to be just the mean of the sample:

Hide

```
(x <- rnorm(5))           # generate a sample of size 5 from the normal
```

```
## [1] -0.01275783 -0.88342562 -0.87266308  0.32030070  0.30106308
```

Hide

```
# fit linear model and get conf. int. of the intercept (sample mean)
confint(lm(x~1))
```

```
##                2.5 %    97.5 %
## (Intercept) -0.9827356 0.5237425
```

Hide

```
summary(lm(x~1))$coefficients # what are the coefficients of the fit?
```

```
##           Estimate Std. Error    t value Pr(>|t|)
## (Intercept) -0.2294966  0.2712962  -0.8459261 0.4452306
```

Hide

```
mean(x)           # just to confirm that intercept of X~1 is indeed the mean:
```

```
## [1] -0.2294966
```

Hide

```
sd(x)
```

```
## [1] 0.6066368
```

Note that the intercept of the $X \sim 1$ model (The ‘Estimate’ column in the summary) is indeed equal to `mean(x)` as the code demonstrates. Now we can use the `mean(x)=x` and `sd(x)=s` in order to compute the 95% confidence interval manually using Eq. (21). The only missing ingredient is z_α , which should be determined, as we discussed, as appropriate quantiles of the t-distribution. Finishing this exercise is left for the homework, just keep in mind that

- for all the distributions available in R, there are $q \dots()$ functions alongside $d \dots()$ and $r \dots()$ (for instance $dnorm()$ is normal distribution probability density, $rnorm()$ is used to draw a random sample from normal distribution, and $qnorm()$ can be used to calculate the quantiles of normal distribution at given percentage cutoffs).

- R does have a built in functions for t-distribution ($>?Distributions$)

- For the sample of size n , its t-statistics follows t-distribution with $n-1$ degrees of freedom.

use $qt()$ to find the t-quantile.

example: Construct a 95% CI for the mean (using one sample).

\bar{x} (sample mean) = 50, s (sample SD) = 10, n (sample size) = 25

Confidence level = 95% $\rightarrow \gamma = 0.95 \rightarrow \alpha = 0.05$, $\therefore t_{0.025}$ for a two-tailed test.

df (degrees of freedom) = $n-1 = 24$

Step 1: find t-quantile ($t_{\alpha, v}$)

$t_{quant} \leftarrow qt(1 - \alpha/2, df) \quad [qt(1 - 0.05/2, 24)] = \sim 2.064$
 \uparrow degrees of freedom (not data frame).

Step 2: find standard error of the mean

$SEM \leftarrow s / \sqrt{n} \quad [10 / \sqrt{25}] = 2$

Step 3: find CIs.

lower CI $\leftarrow \bar{x} - (t_{quant} \times SEM) \quad [50 - (\sim 2.064 \times 2)] = 45.87220 \dots$

upper CI $\leftarrow \bar{x} + (t_{quant} \times SEM) \quad [50 + (\sim 2.064 \times 2)] = 54.127797 \dots$

use $round(lowerCI, 2) = 45.87$

So - with 95% confidence we can estimate that the true population mean falls between 45.87 and 54.13.

Interpretation of the confidence interval:

- In repeated sampling:

If you were to take many samples and calculate this interval each time, 95% of those intervals would contain the true μ .

- For a specific sample:

Once you compute the interval for your sample, it's a fixed range; the true μ either falls within that range or it doesn't. The

"95% confidence" does not mean there's a 95% chance for this specific interval to contain the true μ . Instead it reflects the reliability of the procedure ~~to~~ used to construct the interval.

Connection to hypothesis testing

Think of hypothesis testing - you may reject a hypothesis b/c your \bar{x} falls outside a certain value/region. But CIs give you a continuous view of the uncertainty in your sample mean. Eg: if the interval contains 0, you might say that $\mu=0$ is a plausible hypothesis.

Sample mean

In these examples the confidence interval is built around the \bar{x} . Since \bar{x} is a random variable (varying from sample to sample), if you took another sample, \bar{x} would likely be different, and so would the resulting CI.

Known vs Estimated Var

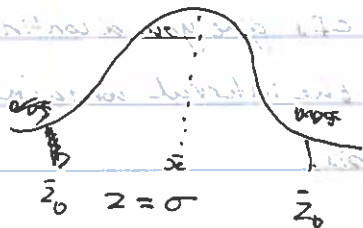
The notes have been working under the assumption that our true σ^2 is known. In most practical situations, you estimate σ from your data, and then typically use a student's t-test to compute the interval instead of the standard normal. This change accounts for the additional uncertainty in estimating the variance.

Equation 8 $P(|Z| > \bar{z}_0) < \alpha$

The probability that a standardized random variable Z (which measures the deviations from its mean in units of standard deviation) exceeds a threshold z_0 in absolute value is less than the significance level α .

↳ i.e. If you calculate how far an observation is from the mean (when scaled by its standard dev) the chance of observing a value even more extreme than threshold z_0 is less than α . For hypothesis testing if this was true ~~we would~~ the observation would be so unlikely + we would accept the null.

Also - this formula is calculating the extreme tail probability.



if $z > \bar{z}_0$ it will be in the far tails.

Good explanation of Interpreting Linear models and confidence intervals

HWQ2 50 points Download a data set from GEO that contains a continuous variable annotation. Build linear models of the effect of gene expression for each gene on the continuous variable (the continuous variable is the dependent variable). Then calculate confidence intervals for each model (use `confint()`, no need to try to calculate the intervals manually). Order the genes by their relationship to the dependent variable (e.g. p-value) and plot the confidence interval values. Describe results in plain English.

The fact that the p-values and coefficients come from a linear model is crucial for interpreting your plot. Let's delve into how this impacts the interpretation:

Linear Model Context:

Linear Model: The linear model you fitted relates gene expression (independent variable) to concentration (dependent variable). The coefficients from this model represent the estimated effect of gene expression on concentration.

P-values: The p-values from the linear model tests the null hypothesis that the coefficient (effect of gene expression on concentration) is zero. A low p-value indicates that the coefficient is significantly different from zero, suggesting a significant relationship between gene expression and concentration.

Relating to Your Plot:

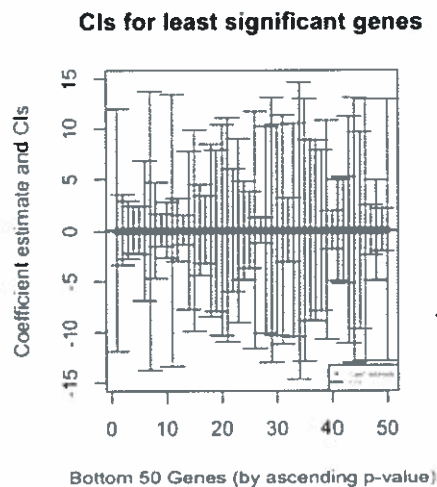
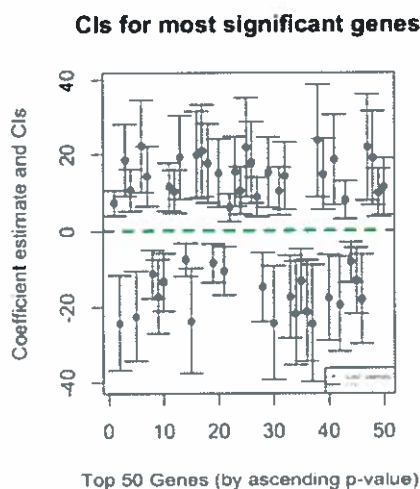
Confidence Intervals (CIs):

The vertical lines represent the range within which the true coefficient (effect of gene expression on concentration) is expected to lie with a certain level of confidence (often 95%). If a CI does not cross the green dashed line at $y=0$, it suggests that the coefficient is significantly different from zero. This indicates a low p-value, suggesting strong evidence against the null hypothesis (that gene expression is not related to concentration).

P-values:

Lower P-values: Genes with lower p-values are likely to be on the left side of your plot, indicating stronger evidence of a significant relationship between gene expression and concentration. These genes have confidence intervals that do not include zero, showing statistical significance.

Higher P-values: Genes with higher p-values are likely to be on the right side of your plot, indicating weaker evidence of a significant relationship. These genes have confidence intervals that include zero, showing no statistical significance.



The coefficient estimate depicts the rate of change of the independent variable.

$$Y = \beta_0 + \beta_1 x + \epsilon$$

Y = dependent variable (concentration)
 β_0 = y-axis intercept \Rightarrow (level of concentration when gene expression = 0).

β_1 = coefficient / rate of change of variable x . As x increases by 1 unit, Y increases by β_1 units.

x = independent variable (gene expression level)

ϵ = error.