Advanced Statistics

author: Bernhard Angele date: Lecture 3

What have we learned last time?

- We figured out that, when we're taking random samples from *any* distribution (with a mean and a variance), the means (and sums) of these samples will approximately follow a **normal distribution** if the sample size is large enough (in general, $n \geq 30$ is a good rule of thumb).
- We even determined what the mean of this sampling distribution of the mean will be (namely, it will be the same as the mean of the population we're sampling from):

$$\mu_{\bar{x}} = \mu$$

- We also determined what the variance of the sampling distribution will be. It will be the variance of the population we're sampling from divided by the sample size:

$$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$$

How can we use this?

- We can do hypothesis testing with normal distributions:
- Let's say we're forensic psychologists trying to screen prisoners for signs of psychopathy.
- Let's say that we have a test that's normed for a standard (prison) population. Thanks to the norming, we know that this standard population has a mean psychopathy score of 50 and a standard deviation of 10. The psychopathy scores (the scores themselves, not just their means) are approximately normally distributed.
- You see a prisoner with a score of 72. Is this an unusually high score? Should you be concerned?

What do you do now?

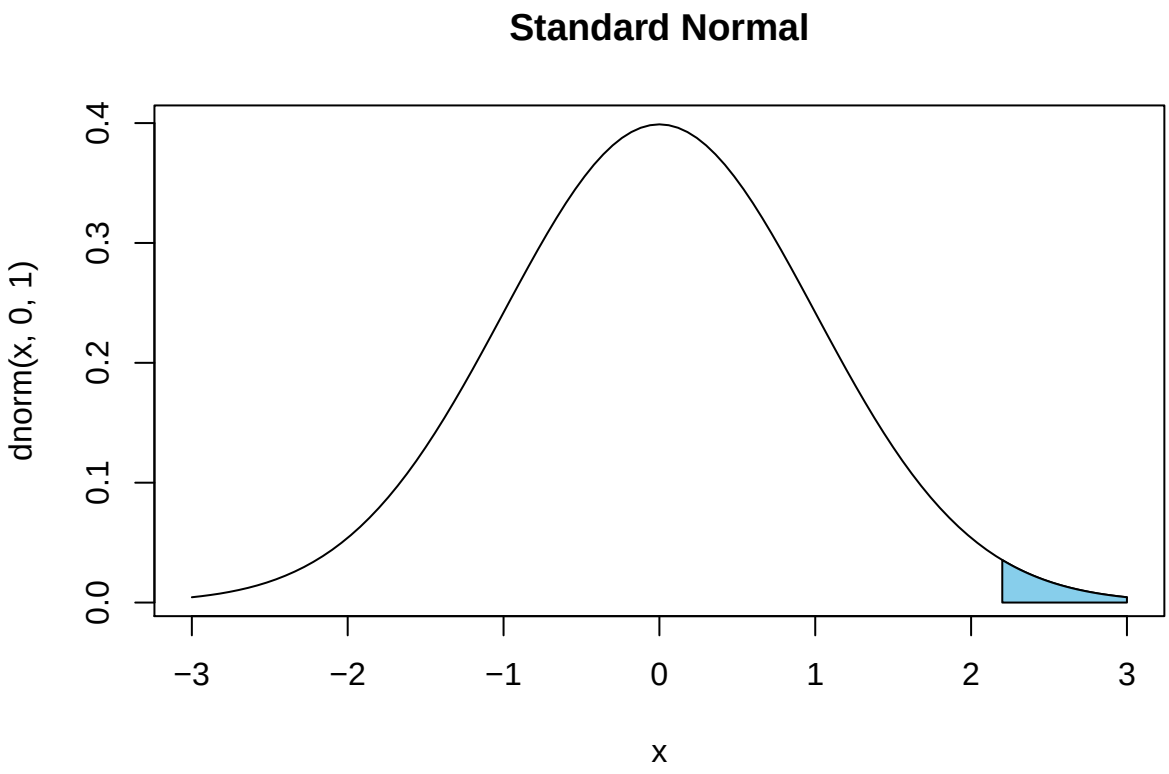
- Establish null and alternative hypotheses:
- Null hypothesis (H_0): the prisoner comes from the standard prison population ($E(x) = \mu_{\bar{x}} = \mu$).
- Alternative hypothesis (H_A): the prisoner's score is higher than that of the standard prison population ($E(x) > \mu$)
- Convert the score into a z-value:

$$z(72) = \frac{72 - 50}{10} = 2.2$$

- What is the probability of getting a z-value of 2.2 given that the null hypothesis is true? - Check the standard normal distribution.

Make a plot

- Always a good idea! A quick sketch is all it takes.



Get the p-value

- Get Excel (or another software) to give you the area under the curve.
- $p(z > 2.2) = 1 - p(z < 2.2)$, so `=1-NORM.S.DIST(2.2,TRUE)`
- Result: 0.0139034
- The prisoner is in the extreme 5% of the distribution.
- Maybe you should be concerned?

Use the EasyStats excel spreadsheet to run many simulations

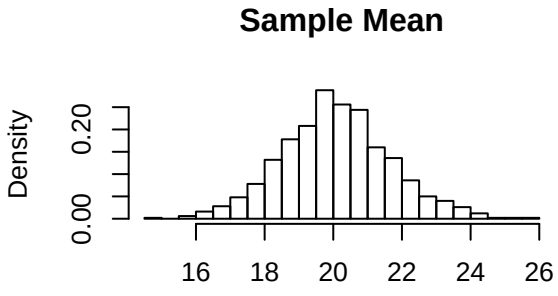
- This is a more intuitive way to do the same thing we did analytically last time.
- Observe:
  - What changes on each run?
  - What stays the same?

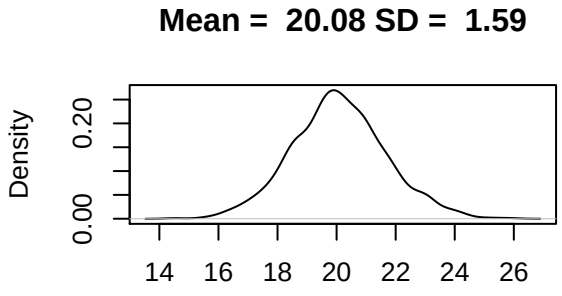
What changes when we re-run the simulation?

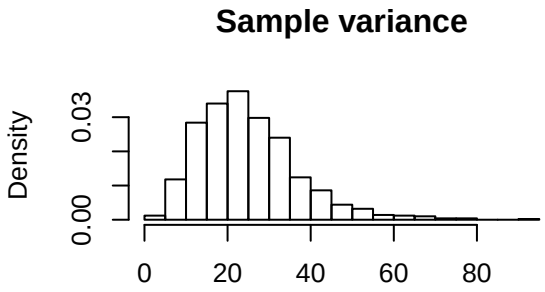
```
## Loading required package: data.table
```

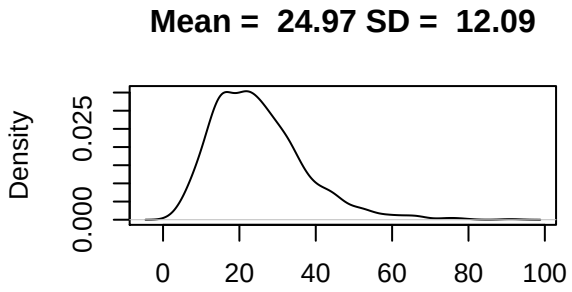
```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
## The following object is masked from 'package:purrr':
##
##     transpose
```

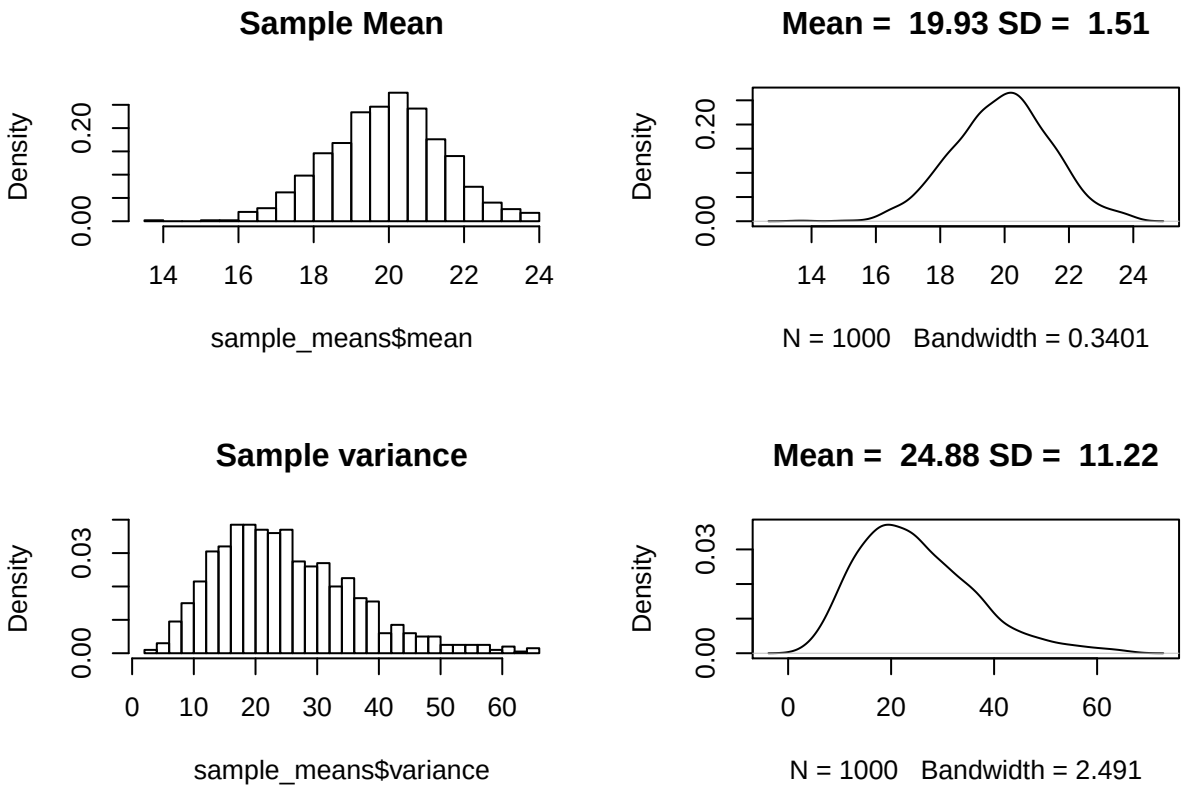








What changes when we re-run the simulation?



What changes when we re-run the simulation?

It turns out the mean of the distribution of sample means varies around the population mean. The sd also varies, but a lot less. It varies around

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

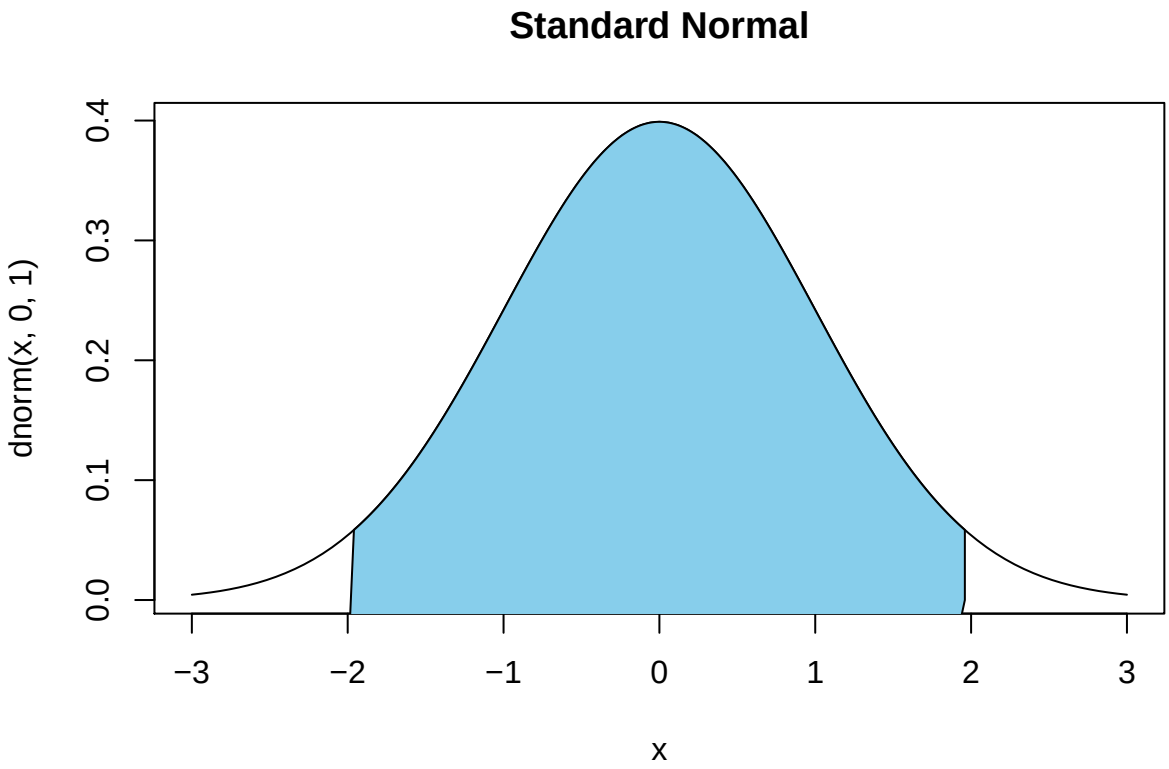
So, to sum up: The distribution of sample means is (roughly) normal, with $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. This means we can apply our knowledge about the normal distribution to find out the theoretical probability of our result given the H_0 (p-values).

Confidence intervals

- A different way of using the normal distribution to express our null hypotheses
- If the distribution of sample means is normal, that means we can say something about the relationship between sample mean and population mean.
- Let's say the population mean μ is 0 and the population sd σ is 1.
- What is the sample mean going to be?
- Think: what is the answer to this going to look like?
- $\mu_{\bar{x}}$ is a random variable, so it doesn't make sense to give a point estimate
- Instead, we can give an interval.

Confidence intervals (2)

- So, let's get the interval that $\mu_{\bar{x}}$ is going to be in 95% of the time.
- We want something like this:



Confidence intervals (3)

- Let's start with the standard normal distribution (**z-scores**)
- We want to get an interval that includes 95% of the area under the curve
- That means we need to take off 2.5% on every side
- For the left interval boundary, we want the x value that is greater than or equal to 2.5% of x values
- ask Excel for the z-value: For this, we use the *inverse* of the standard normal distribution: `=NORM.S.INV(0.025)`
- Result: -1.959964

Confidence intervals (4)

- For the right interval boundary, we want the x value that is greater than or equal to 97.5% of x values.
- ask Excel for the z-value: For this, we use the *inverse* of the standard normal distribution: `=NORM.S.INV(0.975)`
- Result: 1.959964
- If you've done statistics before, these numbers should be pretty familiar to you.
- Generalising this to other normal distributions is easy: $\bar{x} = \mu \pm 1.96 \times \sigma_{\bar{x}}$

- Replacing $\sigma_{\bar{x}}$ with the expression based on the population SD: $\bar{x} = \mu \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$

Exercise

- It is (for some strange reason) well-known that the amount of cat food a cat needs per day is normally distributed with a mean of 2 cans per day and an sd of .5. I'm planning to adopt two (completely random) cats and need to plan this move financially.
- What's the maximum and the minimum amount of cat food cans I must expect to buy every day?
- This estimate should be fairly accurate and should only have a 10% chance of being wrong.
- Suppose I don't care about the minimum amount, I just want to know the maximum – does that change anything?
- Suppose I'm adopting 3 cats instead of 2 – does that change anything about my estimate?

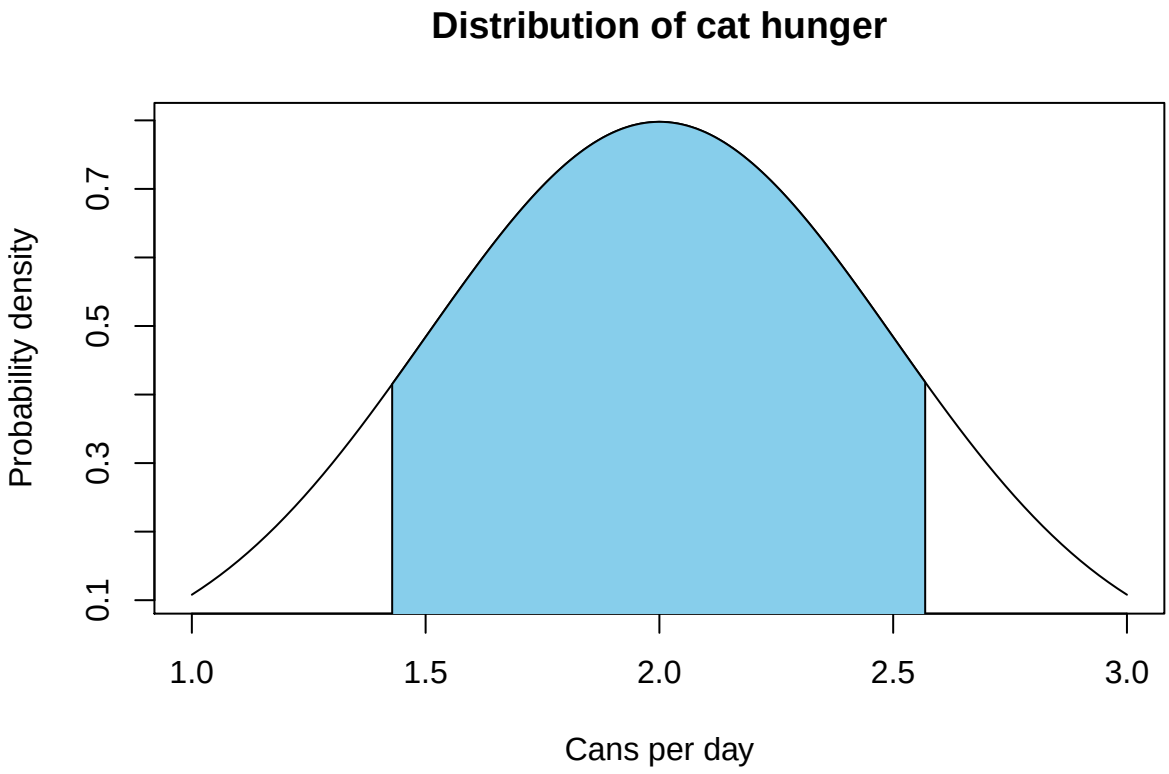
Solution

- I'm drawing a random sample of hungry cats (sample size 2) from the population of hungry cats
- How hungry? Mean = 2 cans/day, sd = .5 cans/day
- I want a 90% CI for the mean of that sample
- Get the z-scores for the lower and the upper bound:
  - lower: `=NORM.S.INV(.05)` = -1.6448536
  - upper: `=NORM.S.INV(.95)` = 1.6448536

Solution (2)

- Calculate the CI:
- lower limit: `=2 + NORM.S.INV(.05) * .5/sqrt(2)` = 1.4184564
- upper limit: `=2 + NORM.S.INV(.95) * .5/sqrt(2)` = 2.5815436
- Those are some hungry cats!
- I need to plan on buying between 1.4184564 and 2.5815436 cans of cat food per day (per cat).

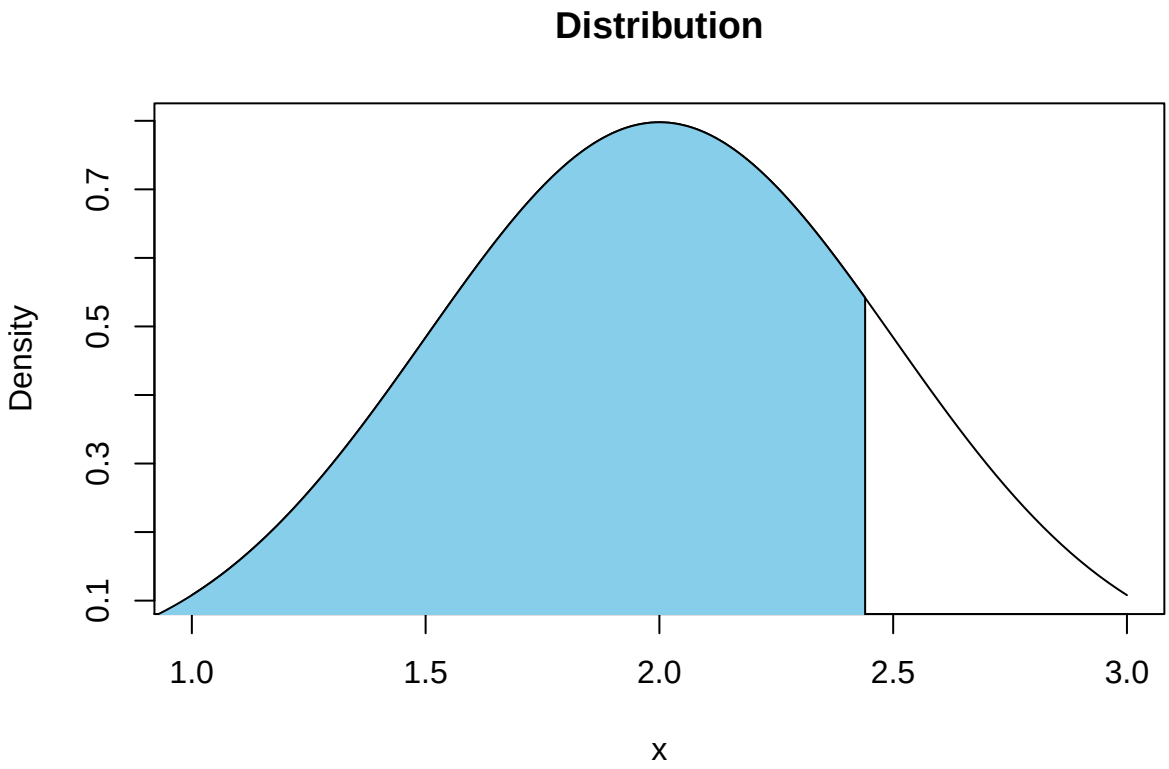
Plot it!



Solution (4)

- If I only care about the maximum, I don't need the lower limit.
- I can use a different upper limit to get an interval that delimits 90% of the area under the curve.
  - upper limit: `=2 + NORM.S.INV(.90) * .5/sqrt(2)`
- Those are still some hungry cats!
- I need to plan on buying at most 2.4530969 cans of cat food per day (per cat).

Plot it again!



Solution (5)

- What if I'm getting 3 cats?
- upper limit: `2 + NORM.S.INV(.90) * .5/sqrt(3)`
- Result: 2.3699521
- Why is it less?
- The chances of getting 3 out of 3 very hungry cats are lower than the chances of getting 2 out of 2 very hungry cats (of course, these figures are per cat, so I'll still have to buy a ridiculous amount of food).

Now reverse the idea

- Usually, we have no other information about a population but the sample we just collected.
- For example, let's say the sample mean is 0 and the sample SD is 1. Apart from this, we know nothing about the population.
- Can we compute a CI for the sample mean?
- Sure enough we can, but it gets a little more complicated.
- (who would have thought?)

What do these numbers mean?

- Anything, really.
- But let's imagine that these numbers are from a survey of student's attitudes towards their Advanced Statistics class.
- Imagine that they could give a rating from -3 ("This is the worst class ever and I want the instructor fired!") to 3 ("This is the best class I've ever taken! I'm going to make so much money with my new R skills!"), with 0 representing a neutral feeling ("It's alright. At least it will be over soon").
- In this case, most students are pretty neutral about the class, but some really love it and some really hate it.

Computing a CI from the sample mean (story time)

- Consider the following scenario:

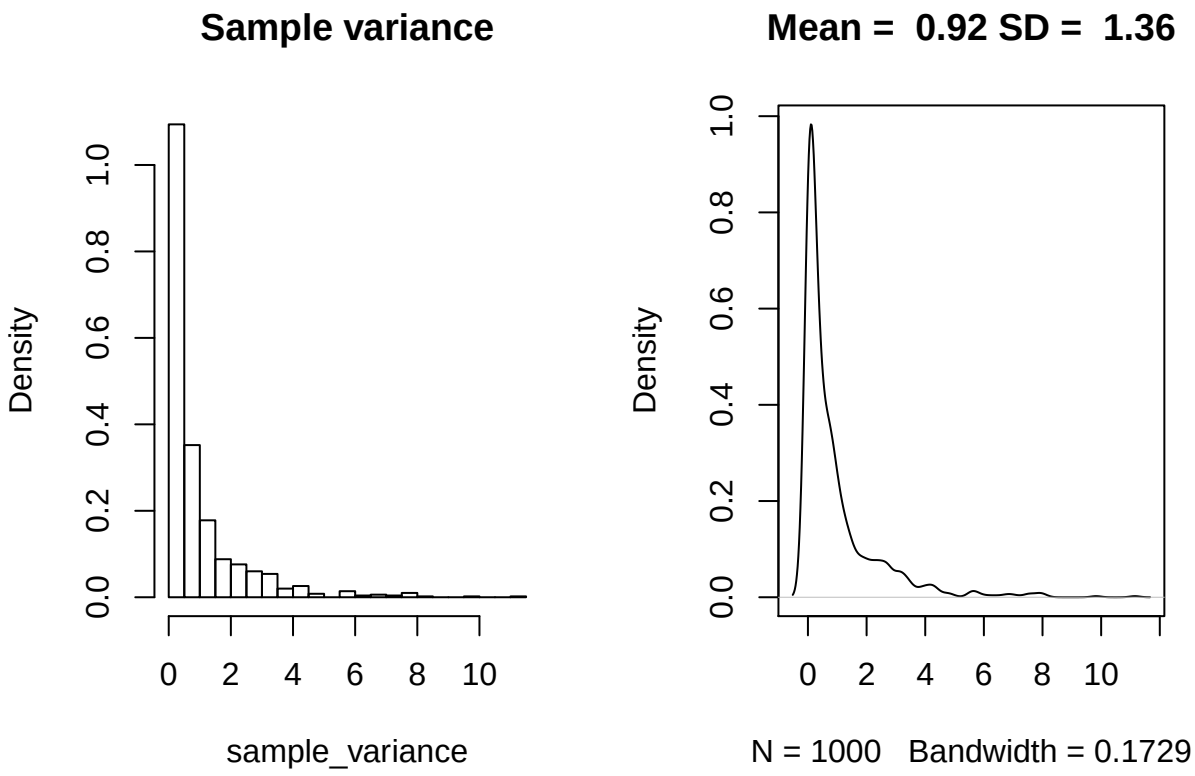
  I have collected 10 responses to my class evaluation (the other students never turned their forms back in). The mean of the responses is 0 (apathy) and the sd is 1. Given that these 10 responses are just a small sample of the population, and that the population I'm really interested in is the population of all current and future Adv Stats students, is there anything I can say about the true population mean? Can I at least conclude that students didn't absolutely hate this class?

Computing a CI from the sample mean

- We'll have to estimate both the population mean and the population variance.
- We have already estabished that the sample mean is a good estimator for the population mean.
- What about the sample sd (s)? Is it a good estimator for the population sd (σ)?
- Or the equivalent question: is sample variance (s^2) a good estimator of population variance (σ^2)?
- We just tackled this in the last lecture analytically
- Today, we can take it easy and just simulate!
- If you haven't done it before, now is a really good time to watch (at the very least) the first 7 minutes of Julian's video (see myBU).

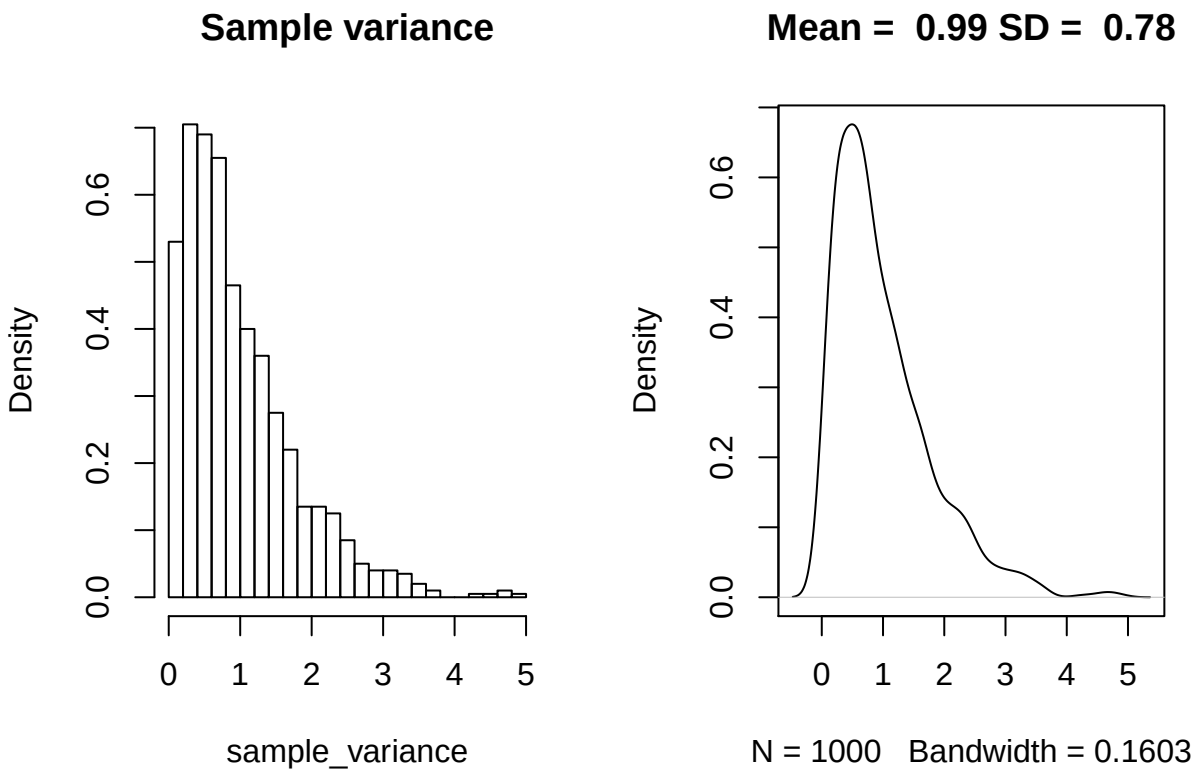
Population variance and sample variance: plots

- Taking samples of size 2 from the standard normal distribution: $X \sim N(0, 1)$ and calculating the variance:



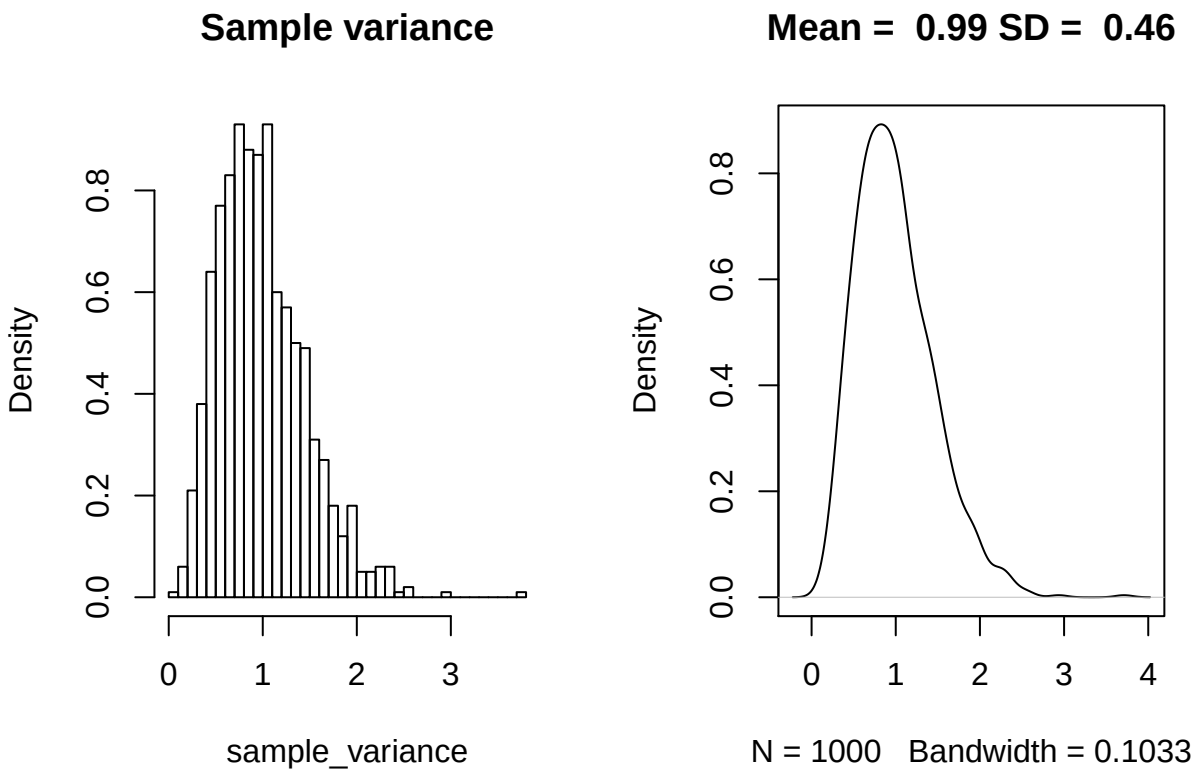
Population variance and sample variance: plots

- Taking samples of size 4 from the standard normal distribution: $X \sim N(0,1)$ and calculating the variance:



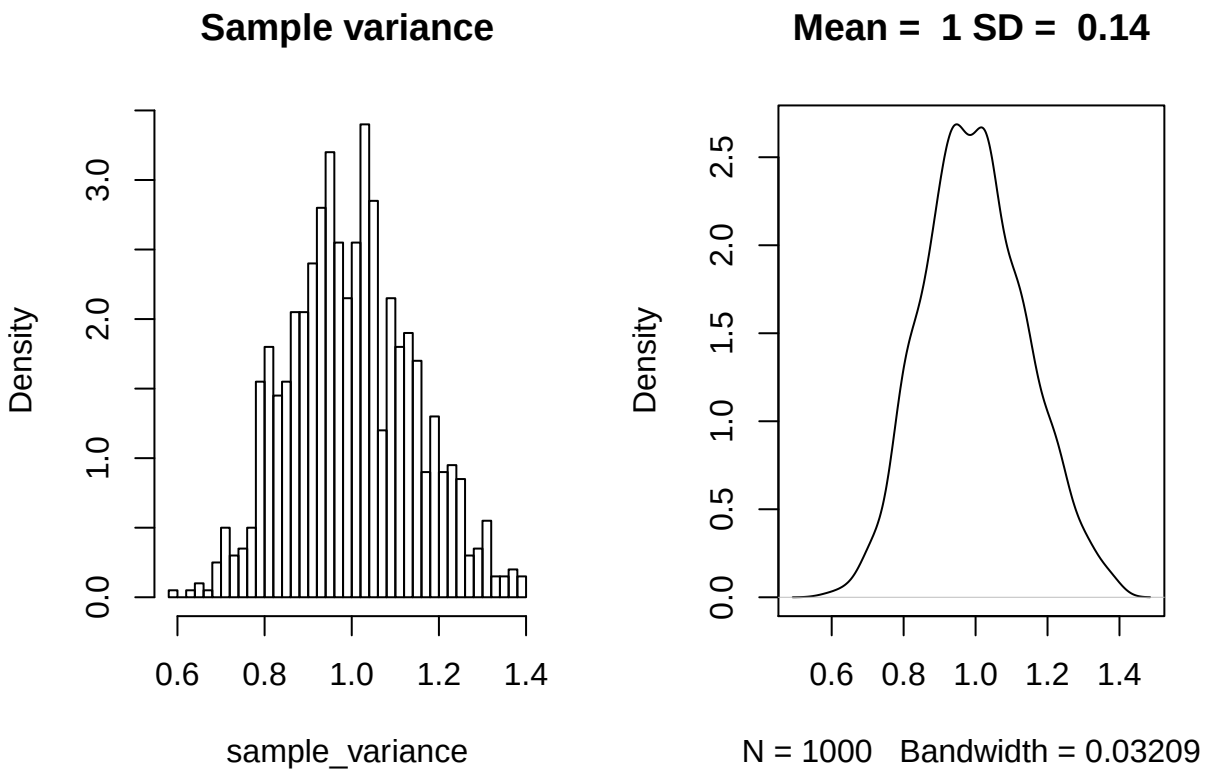
Population variance and sample variance: plots

- Taking samples of size 10 from the standard normal distribution: $X \sim N(0, 1)$ and calculating the variance:



Population variance and sample variance: plots

- Taking samples of size 100 from the standard normal distribution: $X \sim N(0, 1)$ and calculating the variance:



Sample variance as an estimator of population variance

- Looks like sample variance (at least if we calculate it dividing by $n-1$ instead of n is a pretty good estimator (unbiased actually)
- We can plug the sd of the sample into the equation for the SD of the sampling distribution (or rather, the standard error):

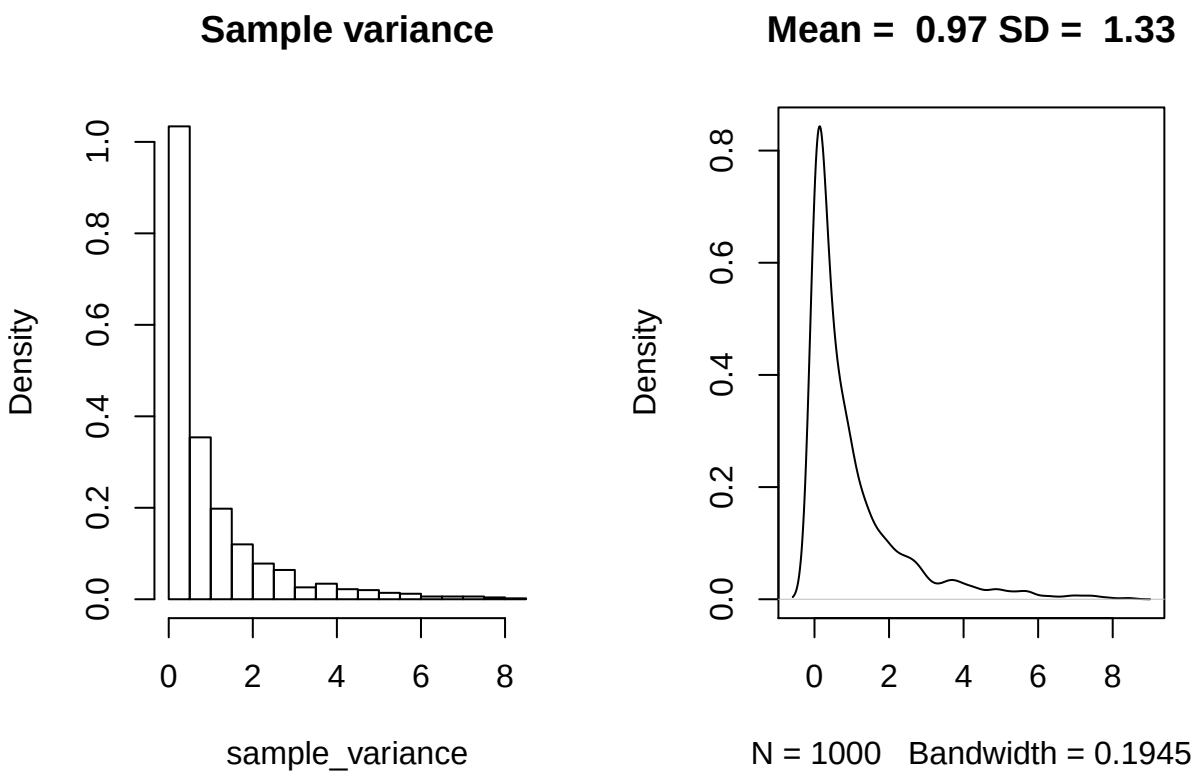
$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

(Note that we are ignoring the question if the relationship between s and s^2 is really the same as the relationship between σ and σ^2. Feel free to simulate that, if you are really curious.)

- But as you saw in Julian's video, the estimate of σ from s is sometimes quite far away from the correct σ, especially for small sample sizes.
- This means that our estimate for σ is going to vary. Its accuracy will depend on the sample size.

The chi-square distribution

- But there's another striking thing going on here. Look again at the distribution of variances for sample size 2:



- This is definitely not a normal distribution!

The chi-square distribution

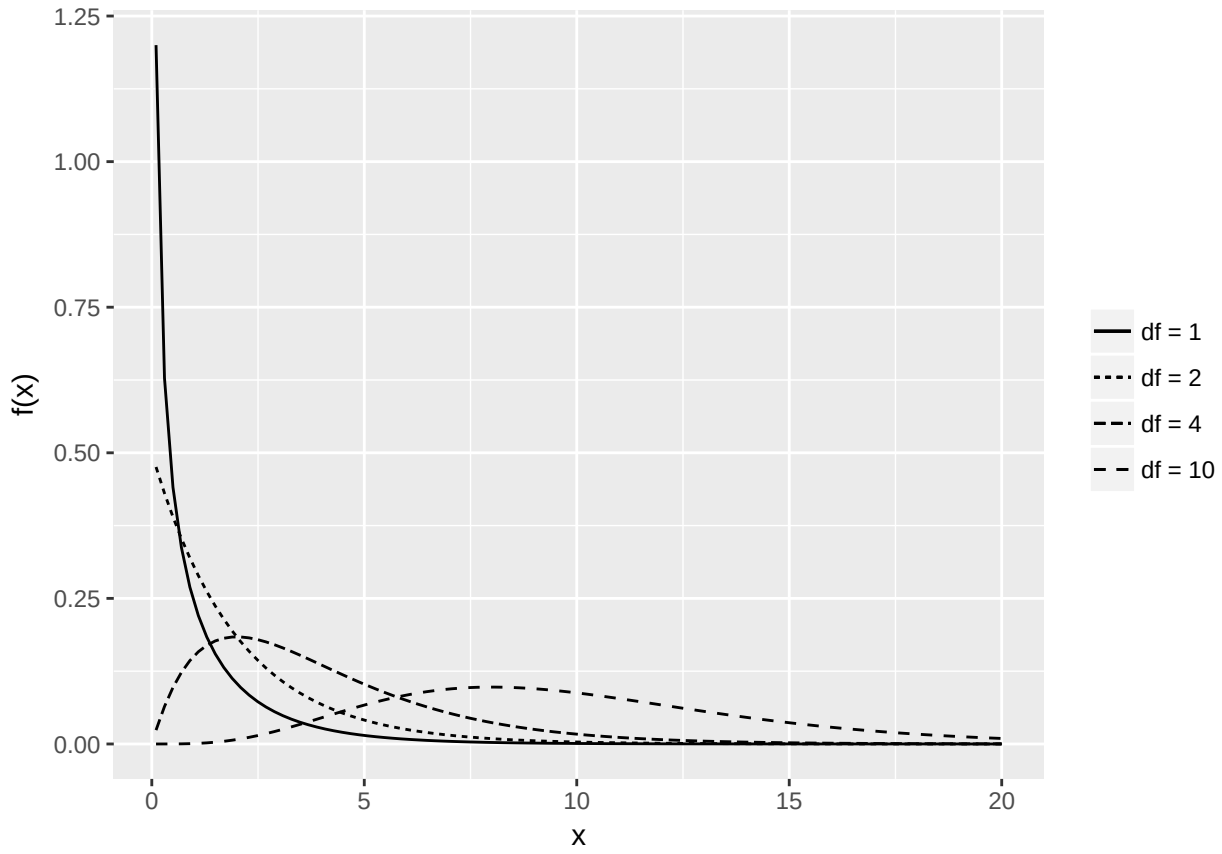
- What is a variance again?
  - Definition: $s^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n-1}$
  - If we have a standard normal distribution ($\mu = 0$ and $\sigma = 1$): $s^2 = \frac{\sum_{i=1}^{n} z_i^2}{n-1}$
  - If $n = 2$: $s^2 = \sum_{i=1}^{n} z_i^2$
  - The χ_1^2 distribution is the distribution of the square of a random variable following the standard normal distribution
    - i.e. squares of z-values are χ_1^2 distributed

Chi-square distributions

- There is more than one χ^2 distribution:
  - The sum of the squares of two **independent**, squared random variables following the standard normal distribution (i.e. z-values) follows the χ_2^2 distribution: $\chi_2^2 = z_1^2 + z_2^2$
  - In general, $\chi_n^2 = \sum_{i=1}^{n} z_i^2$

– Here, n, the number of independent z^2 variables is also known as the **degrees of freedom** of the χ^2 distribution.

Chi-square distributions plotted



What can we do with chi-square?

- We can approach our dice problem in a different way
- Instead of looking at the sample means, we can look at the dice roll results directly
- These come from a distribution called the **multinomial** distribution
- Let's start with coin flips though, because that way we can use the **binomial** distribution

The binomial distribution

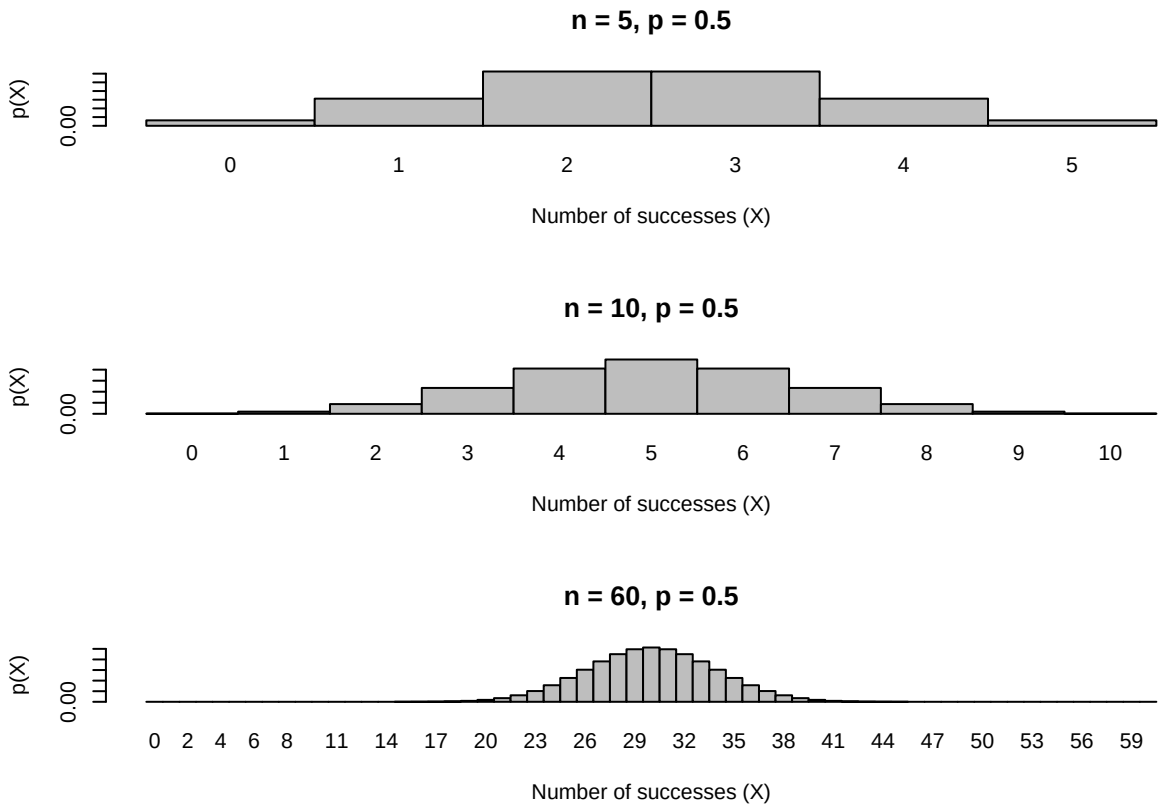
- This is the distribution of number of successes in a sequence of n independent yes/no experiments
- Definition:

$$f(X = k|n, p) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

– Where k is the number of successes (e.g. number of heads), n is the total number of experiments (coin flips), and p is the probability of the success (e.g. 0.5 for a fair coin).

- You almost definitely did this in school, but we won't go into the details of this distribution much. Instead, we'll just look at what happens when we increase the sample size

Plotting the binomial distribution



Binomial and normal distribution

- For large sample sizes, the binomial distribution approximates the normal distribution
- Because of this, there is an easy way of calculating a z-value (or rather, the square of a z-value – I'll spare you the proof, but ask me if you're interested):
  - First, get the frequencies of successes $f_{o(1)}$ and non-successes $f_{o(2)}$.
    - For example, if you had $f_{o(1)} = 40$ times Heads and $f_{o(2)} = 60$ times Tails, can we conclude that the coin is not fair?
  - Then get the expected frequencies given the null hypothesis. If we have a fair coin, our p(Heads) should be .5, so we're expecting $f_{e(1)} = 50$ times Heads and $f_{e(2)} = 50$ times tails.

The chi-square test

- If the sample size is large enough (more than 10 per category), the binomial distribution approximates the normal distribution and the squared differences between the observed ($f_{o(j)}$) and the expected

($f_{e(j)}$) are z^2-values (again, if you want to know why, I can tell you).

$$z^2 = \chi_1^2 = \frac{\sum_{j=1}^{n} (f_{o(j)} - f_{e(j)})^2}{f_{e(j)}}$$

The chi-square test (2)

- In this case, we have two groups (Heads and Tails), so $n = 2$. We can rewrite the sum as:

$$z^2 = \chi_1^2 = \frac{(f_{o(1)} - f_{e(1)})^2}{f_{e(1)}} + \frac{(f_{o(2)} - f_{e(2)})^2}{f_{e(2)}}$$

- Plug in our values ($f_{o(1)} = 40$, $f_{o(2)} = 60$, $f_{e(1)} = f_{e(2)} = 50$):

$$z^2 = \chi_1^2 = \frac{(40-50)^2}{50} + \frac{(60-50)^2}{50} = \frac{100}{50} + \frac{100}{50} = 4$$

- We can look up the probability of getting a value this extreme based on the χ^2-value: `=1-CHISQ.DIST(4,1,TRUE)`, which is 0.0455003
- Conclusion: if the null hypothesis (fair coin, p(H) = .5) is true, we would expect to find an outcome like H: 40, T:60 in less than 5% of samples.

Degrees of freedom

- Wait, what is the `1` in `=1-CHISQ.DIST(4,1,TRUE)`?
  - That's the degrees of freedom. Remember, the degrees of freedom are the number of independent z^2 variables we are summing up.
  - Why only one, when we are summing two terms?

$$z^2 = \chi_1^2 = \frac{(f_{o(1)} - f_{e(1)})^2}{f_{e(1)}} + \frac{(f_{o(2)} - f_{e(2)})^2}{f_{e(2)}}$$

- In this expression, the second term is determined by the first, since $f_{o(2)} = n - f_{o(1)}$.
  - There is only one term that can vary freely, hence $df(\chi^2) = 1$.

Generalising the chi-square test

- Why use χ^2 here at all, when we could just take the square root and do a z-test?
- The answer is that this whole principle generalises to the **multinomial** distribution, i.e. cases where we have more than two groups.
- In the multinomial distribution, we have more than $n = 2$ groups, but the general equation stays the same:

$$\chi_{n-1}^2 = \frac{\sum_{j=1}^{n} (f_{o(j)} - f_{e(j)})^2}{f_{e(j)}}$$

- Our χ^2 is distributed with $n - 1$ degrees of freedom, where n is the group size.

Try it

- The following table is from a dice roll experiment. Use the χ^2 test to decide whether the die was fair or not.

x_i	$f_{o(i)}$
1	18
2	17
3	14
4	19
5	17
6	15

More fun things to do with chi-square

- Remember, we still have the issue of usually not knowing anything at all about the true population mean μ and the true population standard deviation σ
- Instead, we have to estimate them using the sample mean $\bar{x}$ and the sample standard deviation s.
  - Both of this is not a problem at high sample sizes (as you can see very clearly in Julian's video and in our simulations here)
  - But we need a way to account for s being less accurate at low sample sizes.

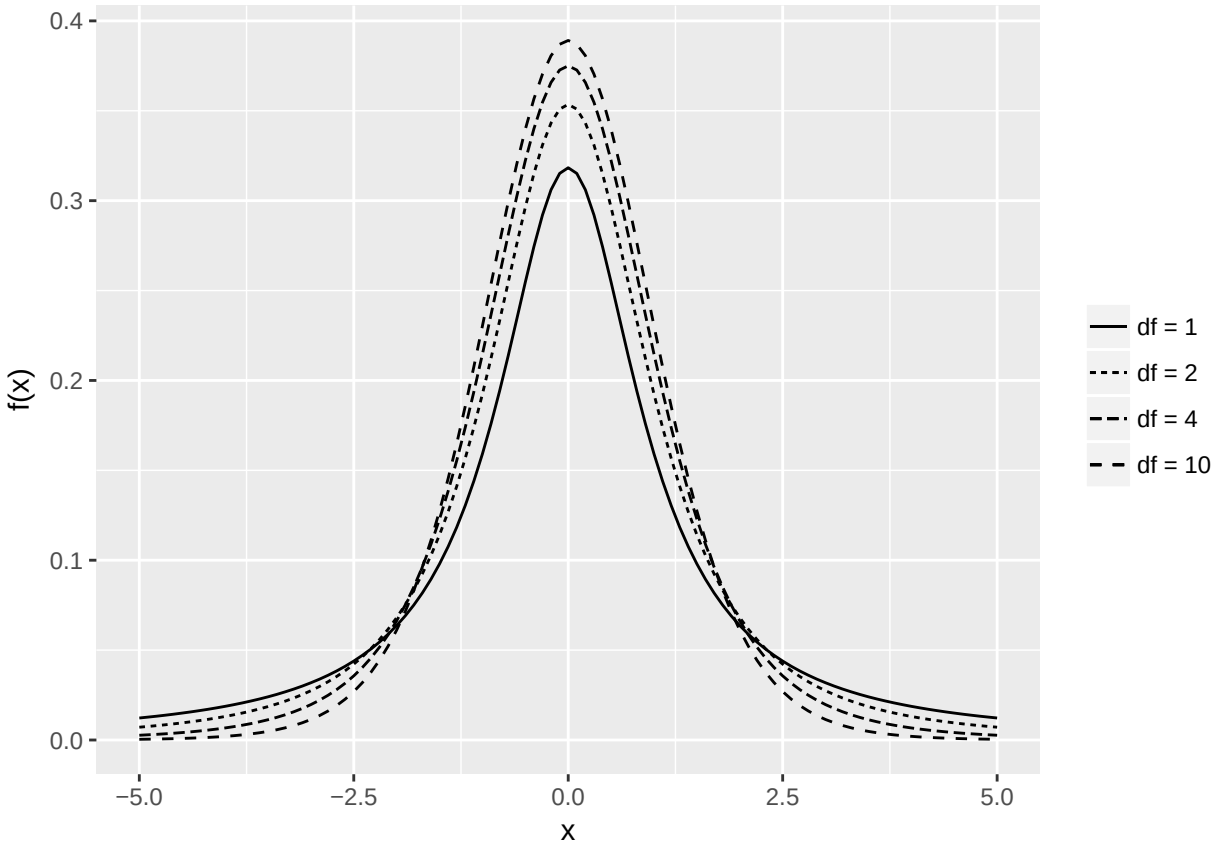
Solution: The t-distribution

- If we divide a z-value by the square root of an *independent* χ^2 value divided by n, we get a t-value:

$$t_n = \frac{z}{\sqrt{\chi_n^2/n}}$$

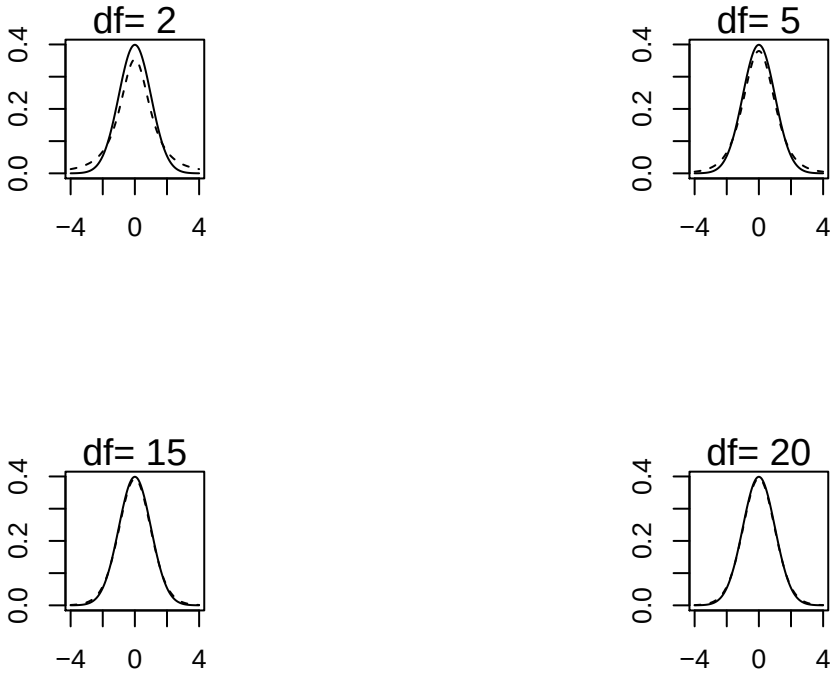
- The t-value has **degrees of freedom** as well – it inherits them from the χ^2 value in its denominator.
- Practically, the denominator makes the distribution have "heavier" tails – exactly what we need for our problem.

Let's plot some t-distributions



The t-distribution vs. the normal distribution

- Solid = normal distribution, dashed = t-distribution



The t-test

- Solution: assume that the sample means aren't normally distributed, but rather t-distributed
- Why t?
- The t-distribution is like the standard normal distribution, but it has an additional parameter that we call df (for degrees of freedom, but don't worry about the name yet).
- The higher df, the closer the t-distribution is to the standard normal distribution
- For lower df, the t-distribution has "heavy tails", meaning that it's wider
  - This reflects greater uncertainty.

t as a test statistic

- Once again, the mathematical proof of this would take too long, but you can show that

$$t_{n-1} = \frac{\bar{x} - \mu_0}{\hat{\sigma}_{\bar{x}}}$$

, where μ_0 is the mean according to the null hypothesis and $\hat{\sigma}_{\bar{x}}$ is the estimate of the standard error of the mean (based on the sample standard deviation) is t-distributed with $df = n - 1$ degrees of freedom.

Back to my little example

- Consider the following scenario:

I have collected 10 responses to my class evaluation (the other students never turned their forms back in). The mean of the responses is 0 (apathy) and the sd is 1. Given that these 10 responses are just a small sample of the population, and that the population I'm really interested in is the population of all current and future Adv Stats students, is there anything I can say about the true population mean? Can I at least conclude that students didn't absolutely hate this class?

Computing the 95% CI

- Using the t-distribution, we can compute CIs from samples as follows: get the lower and upper bounds from the t-distribution (which one depends on the sample size, e.g. in this we have $n = 10$, so we will use a t-distribution with $df = n - 1 = 9$):
- Lower bound (remember, we want to exclude the extreme low 2.5%): `=T.INV(0.025, 9)` (where 9 is the df)
  - Result: -2.2621572
- Upper bound (remember, we want to exclude the extreme high 2.5%): `=T.INV(0.975, 9)` (where 9 is the df)
  - Result: 2.2621572
- No surprise: the t-distribution is symmetrical

Computing CIs

- Then take the upper and lower bounds and compute the CIs as follows: $\bar{x} = \mu_{\bar{x}} \pm 2.262 \cdot \frac{s}{\sqrt{n}}$
- Remember, we estimated $\mu_{\bar{x}}$ using the sample mean (in our example, $\bar{x} = 0$) and the population variance σ using the sample standard deviation (in our example, $s = 1$).
- CI: $0 \pm 2.262 \cdot \frac{1}{\sqrt{10}} = 0 \pm .7153$
- Lower bound: -.7153
- Upper bound: .7153

Back to our example

- Hey, there is a good chance that my current and future students don't absolutely hate me (yet)!
- The lowest mean in the CI is -.7153, which maybe translates to "apathetic but slightly worried."
- But they don't love me either:
- The highest mean in the CI is .7153, which maybe translates to "apathetic but slightly hopeful."
- Of course, the true mean is actually outside 5% of the intervals calculated like this.

What does the CI of the sample mean mean? (sorry)

- Remember, we are reversing the idea that the sample mean has a 95% probability to be within the 95% confidence interval around the population mean.
- When we calculate a 95% CI from a *sample* this **DOES NOT MEAN** that there is a 95% probability that the population mean is within this 95% CI.
- The true mean either is or is not in this particular CI.
- Rather, it means that if you take a lot of samples and compute the CI around the sample mean, 95% of those CIs will contain the true population mean.
- In other words, the CI bounds are random variables, but the population mean isn't.

- (In Bayesian statistics, you can actually get something equivalent to the first definition – a 95% credible interval.)

Let's test this

- Let's get 10 samples from a normal distribution, then get CIs from them and see how often they contain the true mean.
- We'll do this in class.
- Spoiler:
- The proportion of CIs that does not contain the true mean is larger than 5%! This is because the normal distribution is narrower than the t-distribution at low dfs.
- Be **very** careful! If you *think* you have a 95% CI, but you actually have a 90% CI or worse, you are prone to making errors in interpreting the results.
  - Horrible, money-wasting, science-distorting, extermely expensive errors!

Hypothesis tests

- In a way, by calculating the CI we already have a way to test hypotheses
- Let's say we got a 95% CI from our sample with a lower bound of 2 and an upper bound of 3.
- Let's use the simplest null hypothesis possible
- Null hypothesis: the mean of the population that the sample came from is 0
- $H_0 : \mu = 0$
- Given the 95% CI above, can we reject the null hypothesis?
  - And if so, what is the chance that we're wrong?
- Answer: Yes, we can, since 0 is not part of the CI.
  - There is the possibility that we are wrong, though, since only 95% of the CIs will contain the true population mean.
  - This is called the type I error, and its probability here (called α)is 5%.

Example

- Remember my survey? The CI did not contain -3, so I can conclude (with an α of 5%), that the average member of the population of current and future Adv Stats students attitude towards me is not intense hatred. Relief!

Two-tailed t-tests

- Instead of computing the CI from the t-value, we can also just take the t-value itself as a measure of how far the sample mean is away from the mean specified in the null hypothesis.
- We can determine a critical t-value t_{crit} depending on our α criterion and the df. For example, for a df of 9, t_{crit} for the upper bound is `=T.INV(.975,9)`, which gives us2.2621572 and t_{crit} for the lower bound is `=T.INV(.025,9)`, which gives us-2.2621572
- Note that the t-distribution is symmetrical In short, if $t \geq |t_{crit}|$, we can reject the null hypothesis.

What about one-tailed t-tests?

- If we are absolutely sure of the direction of the effect, then we could use a t-test that only rejects the null hypothesis when the t-value is greater than t_{crit} or if it is smaller than t_{crit} (depending on what direction we want to test for).
- In this case, our t_{crit} can be a little closer to 0, since the entire 5% rejection area is in one tail only: `=T.INV(.95,9)`, which gives us 1.8331129.
- But be careful, if the effect is in the wrong direction (even if it's ridiculously strong in the wrong direction), we can't reject the null hypothesis with that test.
- This is one of the weird cases in null hypothesis significance testing (NHST) where our intentions can determine the results of the test. Bayesian statisticians are right to complain about this.

Example

> I'm trying a new type of medication to help insomniac patients sleep better. Each of my 5 patients reports how much longer (or shorter) they have been sleeping (in hours) after taking the medication compared to before. The numbers are below. Based on this, can I conclude that the medication has changed my patients' sleep? Or are the variations that the patients observed random and unrelated to the intervention?

```
## [1]  2.35 -0.22  2.44  0.19  1.43
```

Your turn. What is the null hypothesis?

Example solution

The H_0 is that the true mean of the population is 0.

```
t.test(sleep_times)
```

```
##
##  One Sample t-test
##
## data:  sleep_times
## t = 2.2712, df = 4, p-value = 0.08561
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.2753893  2.7513893
## sample estimates:
## mean of x
##     1.238
```

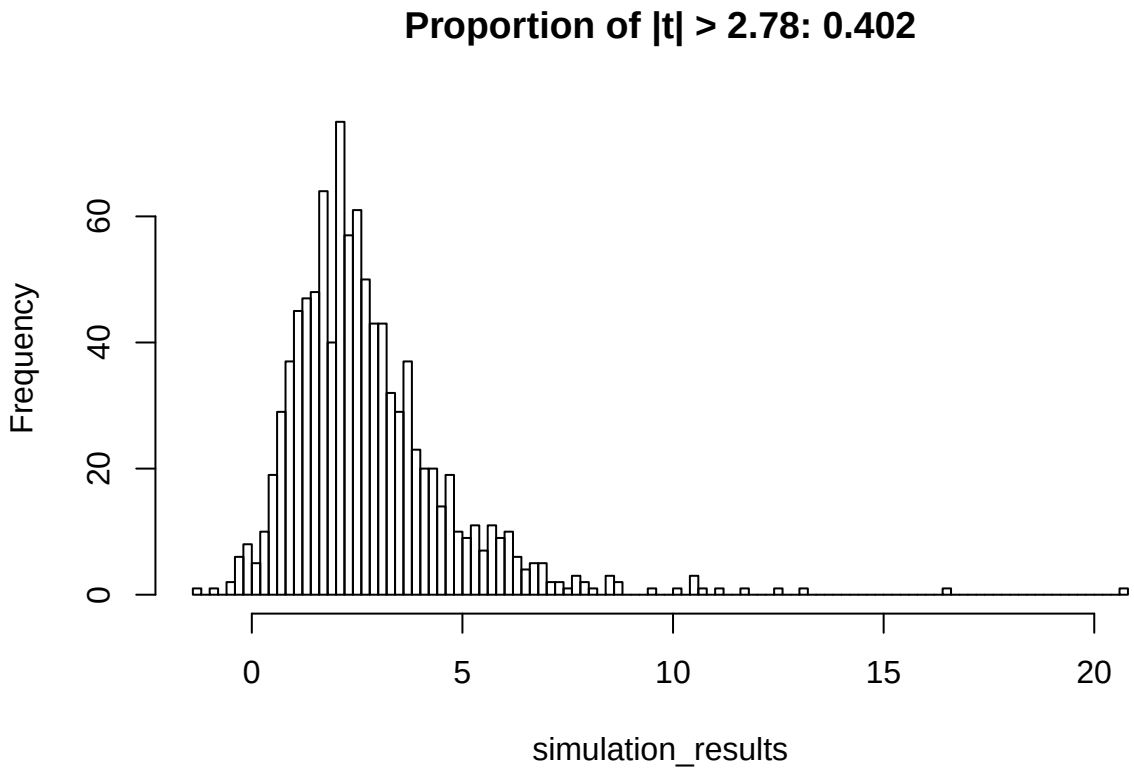
-If p $\leq$ .05: reject the null hypothesis. - Try this in SPSS!

Power simulations

- For simple (and even more complex) designs, you can compute power analytically. I will show you how to do this using a program called GPower.
- But simulations are a lot more intuitive!
- Let's go back to the sleep example and assume that the true mean was 1 (that means that on average, people get one hour more sleep when using the medication) and the sd was 1.

Power simulations plot

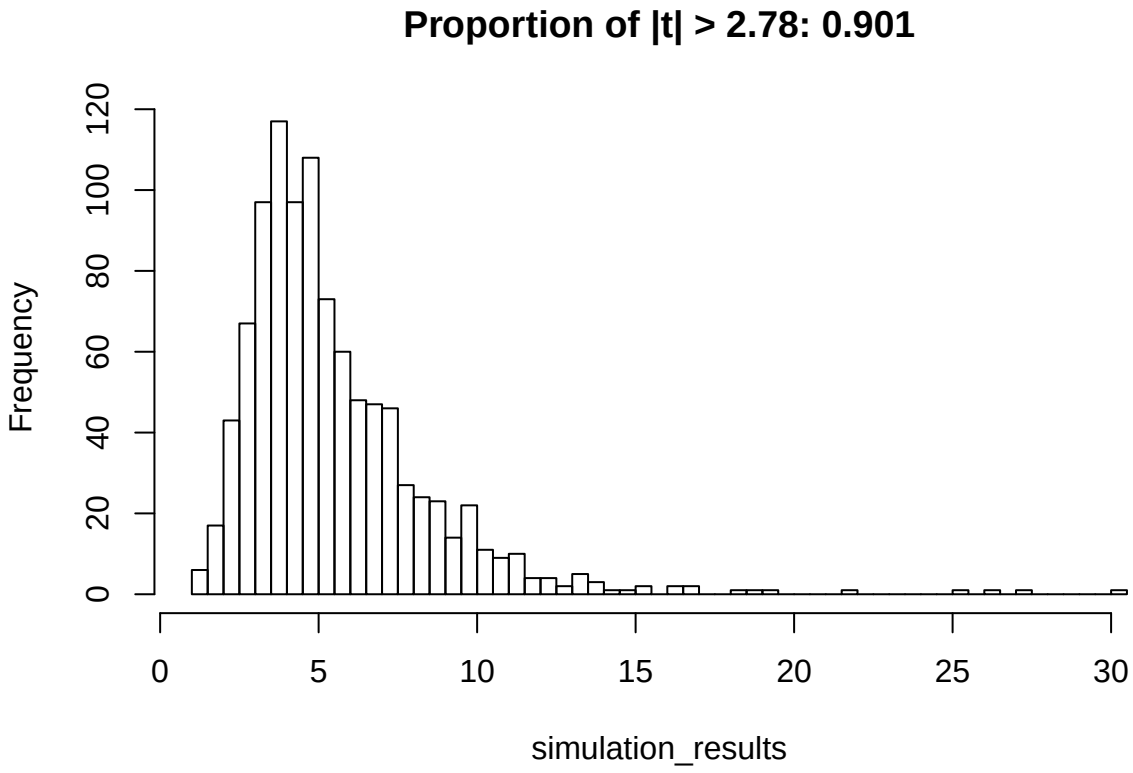
- Remember, $t_{crit} = 2.7764451$



Not so great!

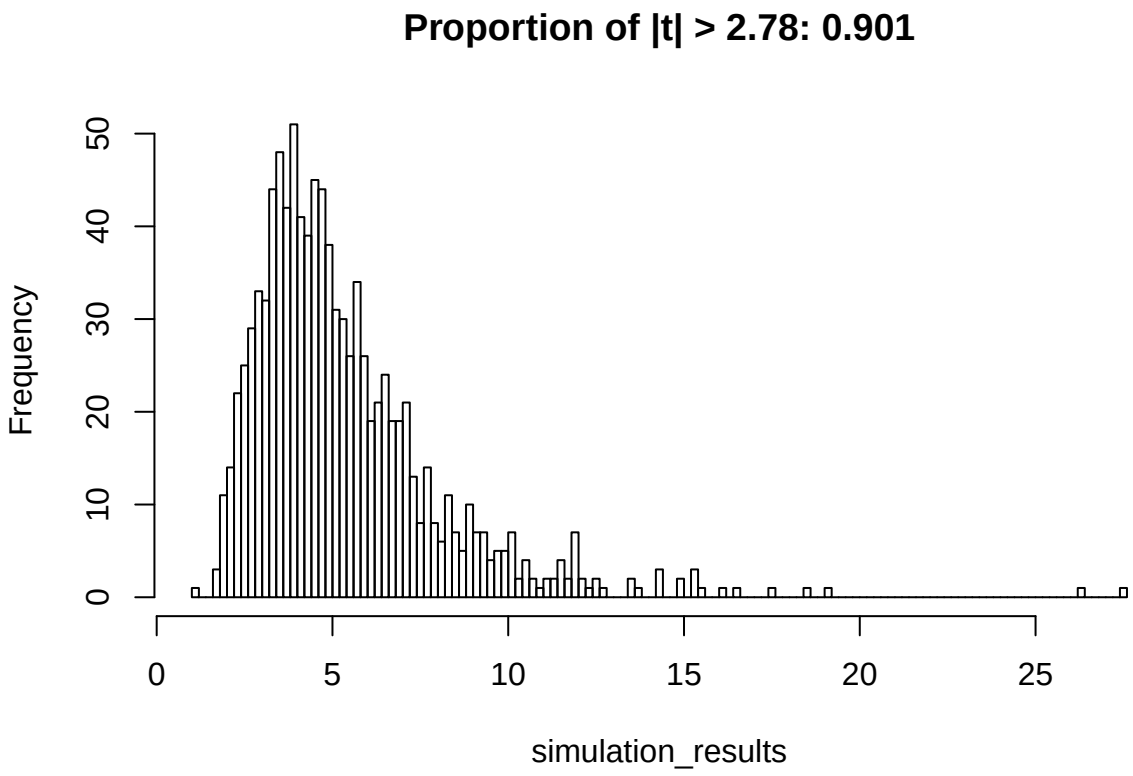
How to increase power

- Let's try a higher true mean ($\mu = 2$):



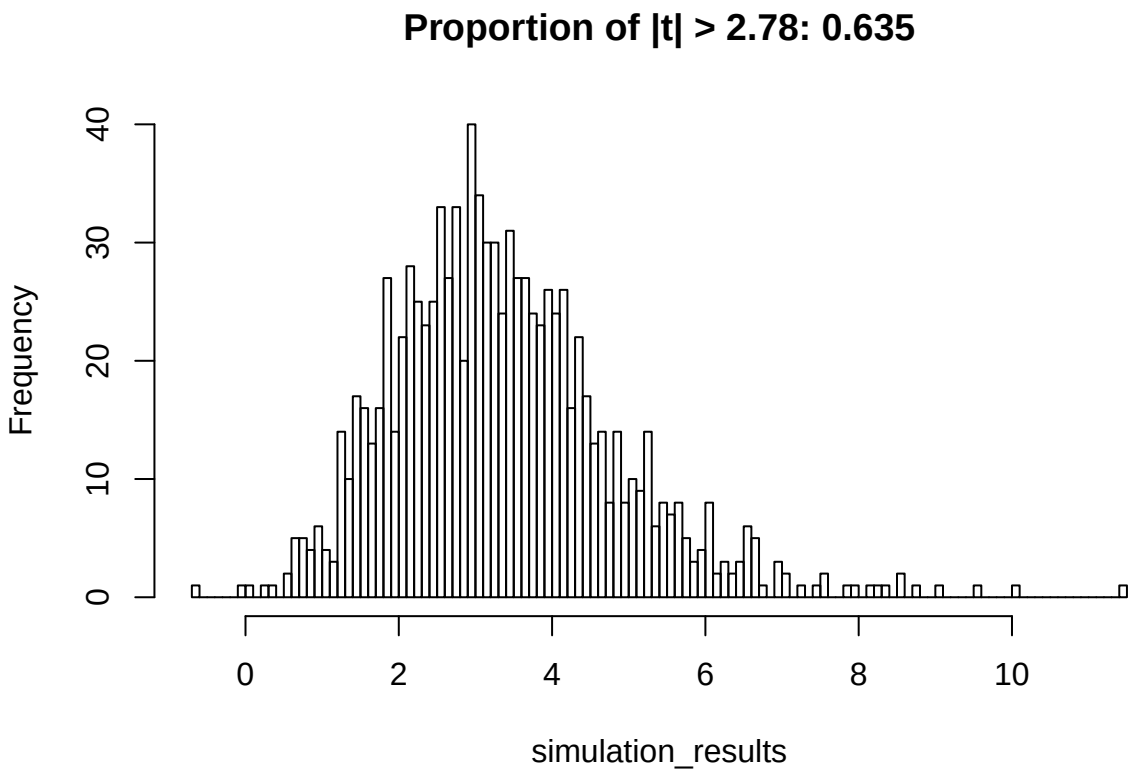
How to increase power (2)

- The standard deviation (i.e. the noise) in the population is lower (people don't vary as much in their response to the medication)
- Let's try a true value of $\sigma = 0.5$



How to increase power (realistically!)

- You don't really have any direct control over population mean (i.e. effect size) or sd (i.e. noise). Let's focus on the one variable that you do have control over.
- The sample size is larger: let's try $n = 10$



Double-checking our results

- Let's just check analytically that we have this correctly: If we want to show in the one-sample t-test that a mean of 1 is different from 0 (when $sd = 1$), we need about 10 subjects. I will show you how to use GPower for that.

Setting yourself up for success (or failure)

- You don't want to run an underpowered study. Most likely, you'll get a null result that tells you nothing about the true state of the world.
- How can you avoid this?
- Run a realistic number of participants so you reach acceptable power (the APA recomments .8).

Exercise

An experimenter knows for a fact that the average number of friends people have on Facebook is 70, with an sd of 10. She knows this because she works for Facebook and has access to all your personal data. The experimenter wants to know if people who post lots of photos of cats have more or fewer friends than the average Facebook user. Automatically tagging cat photos is hard, so our experimenter just asks an unpaid intern to compile a sample of 100 cat-posting people and

find out their friend numbers. How big does the effect (in friends gained/lost) have to be so it would be detectable at an acceptable power level of .8?

Solution

- Effect size is defined as
$$d = \frac{\mu_1 - \mu_2}{\hat{\sigma}}$$
- We're doing a t-test where we want to know if the group mean (for a group size of $n = 100$) comes from a known population ($\mu = 70, \sigma = 10$)
- We want to find the necessary effect size given the sample size and the standard error of the mean. The test would be two-tailed, since we don't know the direction of the effect. The power we want is $(1 - \beta) = .8$

In GPower

- In GPower, select `t tests` as `Test family` and `Means: Difference from constant (one sample case)` as `Statistical test`. As `Type of power analysis`, select `Sensitivity: Compute required effect size`
- In the `Input Parameters` area, select `Two` for `Tails`, leave the α at `.05`, set the `Power` to `0.8`, and set the `Total sample size` to `100`. You get an effect size of $d = 0.2829125$
- A difference in as little as $d \cdot \sigma = .283 \cdot 10 = 2.83$ friends would be detectable.

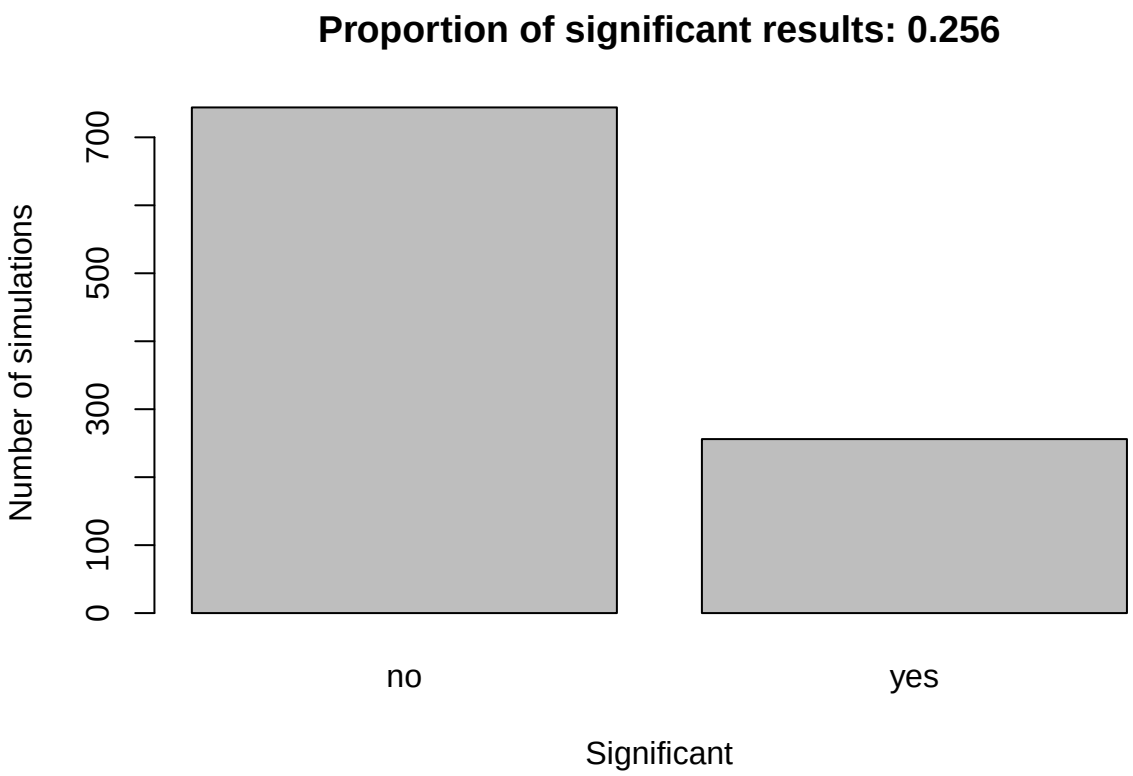
Don't cheat!

- How about the following strategy?

  Just run the hypothesis test on the data after every new sample and stop as soon as you get a significant result.
- Let's see just what happens to α if you do that.
- Run a simulation where there is no effect (i.e. where we know the H_0 is true)

The consequences of cheating

Let's run this simulation 1000 times and make a plot with the results:



The consequences of cheating (2)

- Holy inflated Type I error rate, Batman!
- α is at 25%, instead of 5% where it should be.
- Unfortunately, this strategy of using stopping rules ("data peeking") is quite common.
- Solution: do a power analysis, set your sample size beforehand, and stick to it!

Testing more interesting hypotheses

- So far, we have been testing the null hypothesis that our sample mean is 0.
- This is not what we usually do in Psychology.
- Instead, we want to know if there is a significant difference between the means of two (or more) samples.
- For example, you might give only one group an intervention against anxiety, with the other one serving as the control.
  - Does the intervention work?
  - Do people in the treatment group report lower anxiety?
  - Can we generalise this to the population?
  - Should we use this intervention in clinical practice?
- A lot of effort and money may be wasted if you get these questions wrong.

The two-sample t-test

- Remember, we are comparing two samples now. We'll call the sample means μ_1 and μ_2.
- Our null hypothesis is $H_0 : \mu_1 = \mu_2$
- We can rephrase this as $H_0 : \mu_1 - \mu_2 = \delta = 0$
- We already know the logic of this: we just want to find out if d (δ = true population difference, d = sample difference)is extreme enough so we can reject the H_0.
- Let's see how δ is distributed.
- We could do this analytically, using the things we've learned about expected values, but I'll leave that to those of you who are really really interested and just give you the end results.

The two-sample t-test (4)

- We're looking for the distribution of sample differences $x_1 - x_2$:
- If we knew the population standard deviation, we could use the following formulas:

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$$

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- Based on this, we could calculate CIs or just a z-value
- Remember our $H_0 : \mu_1 - \mu_2 = 0$
- The z-value would then be: $z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$
- (since our null hypothesis is that $\mu_1 - \mu_2 = 0$)

The two-sample t-test (5)

- Of course, in real life we don't know the population sd
- So we have to estimate it using s^2
- This would be a t-value, not a z-value

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- Only problem: what is the df of that test? If the variances are equal, it's easy: $df = n_1 + n_2 - 2$
- There are some shortcuts that we can take if the sample sizes and population variances are the same, but is there a general solution?
- We could just use the lower of the sample sizes, but this will cost us power
- This was actually a big problem in statistics, but B. L. Welch found an approximate solution (called *Welch's t-test*)
- You can look the details up on Wikipedia, but SPSS knows them and will apply them automatically.

The dependent t-test for paired samples

- This is actually a lot easier. Since we have two samples per person/group/analysis unit, we can simply compute the differences between measurements and then use the one-sample t-test to check if they are 0.

- First, we calculate the mean and the standard deviation of our sample of n difference values $d_i = x_{i1} - x_{i2}$:

$$\hat{\mu}_d = \bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$$

$$\hat{\sigma}_d = s_d = \sqrt{\frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n-1}}$$

$$\hat{\sigma}_{\bar{d}} = \frac{\hat{\sigma}_d}{\sqrt{n}} = \frac{\hat{s}_d}{\sqrt{n}} = \frac{\sqrt{\frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n-1}}}{\sqrt{n}}$$

- Here, n is the number of sample *pairs*

The dependent t-test for paired samples (2)

- Then we can calculate the t-value:

$$t = \frac{\bar{d} - \mu_d}{\hat{\sigma}_d}$$

, where μ_d is the population mean for the difference given that the H_0 is true. If the H_0 is that both samples are the same ($\mu_d = 0$), this simplifies to

$$t = \frac{\bar{d}}{\hat{\sigma}_d}$$

-We can estimate the standard error of the difference mean $\hat{\sigma}_{\bar{d}}$ from the standard deviation of the difference:

$$\hat{\sigma}_{\bar{d}} = \frac{s_d}{\sqrt{n}}$$

-Plugging this into the equation for t, we get:

$$t_{n-1} = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$$

- The resulting t-value will have a df of $n - 1$, where n is the number of sample pairs.
- Since the sd of the differences will be a lot lower than the overall sd, the power of this test is quite a bit higher.