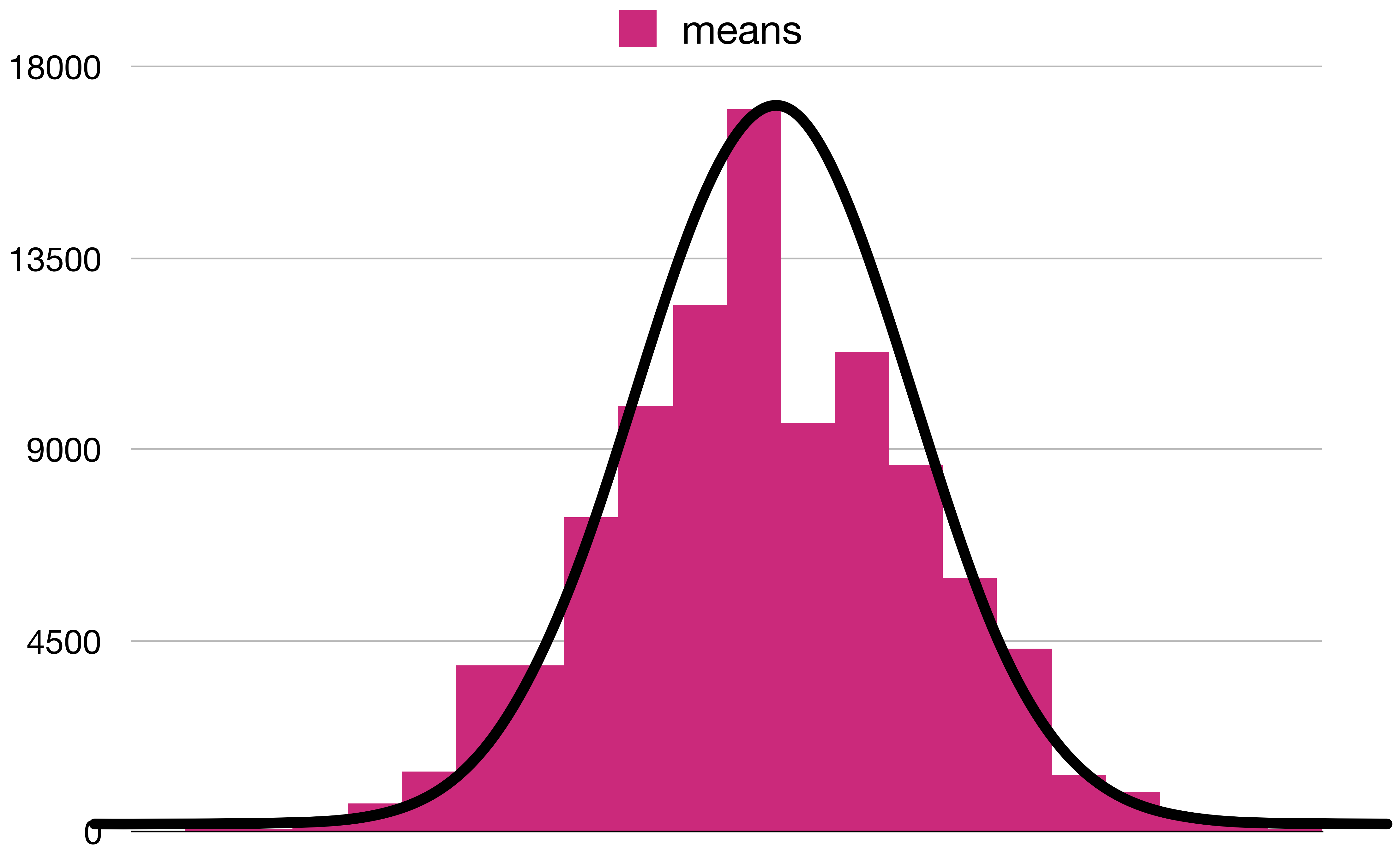
ECE 20875
Python for Data Science

Milind Kulkarni and Chris Brinton

confidence intervals and hypothesis testing

sampling distribution



Each data point is the $\bar{x}$ of one experiment

- Recall that by the central limit theorem, sample means approach a normal distribution
- Can we use this to draw conclusions about our data?

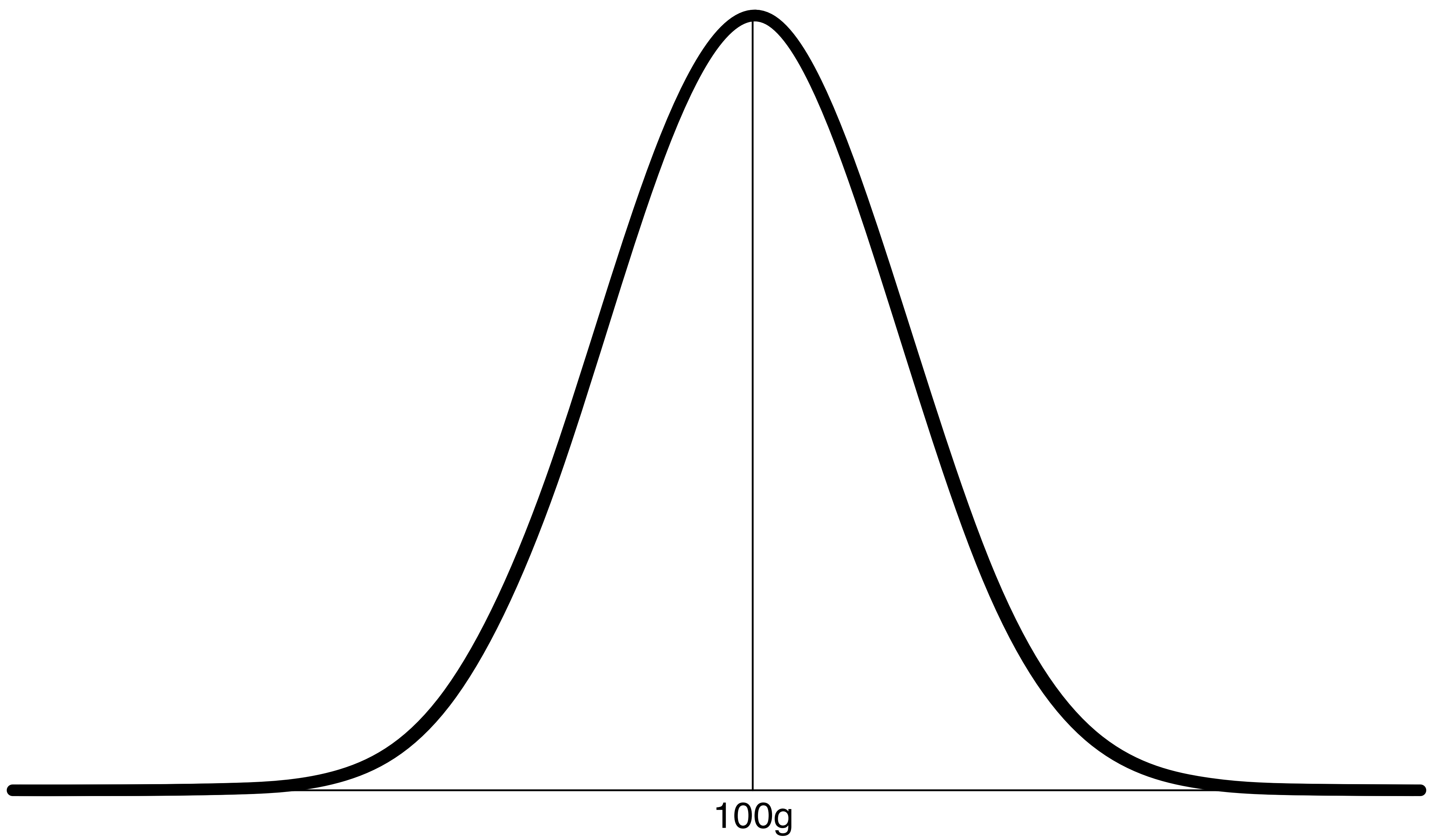
asking questions about data

- Suppose a factory claims to produce widgets with an average weight of 100g and a standard deviation of 22g
- We receive a new shipment of widgets which seem off, and we want to see whether the factory has shifted
- Form two hypotheses:
  - **Null hypothesis** (H_0): The factory is producing according to specification, i.e., $\mu = 100g$.
  - **Alternative hypothesis** (H_1): The factory is not producing according to specification, i.e., $\mu \neq 100g$.
- Suppose we weigh 100 of the new widgets (i.e., sample $n = 100$ widgets) and find their average weight is $\bar{x} = 95g$
  - What can we conclude?

asking questions about data

- **Are the widgets in spec?**
- Not as simple as it seems!
- We have picked one sample of widgets, but it could just be a bad sample!
- Can we use our sampling distribution to help?

hypothesis testing

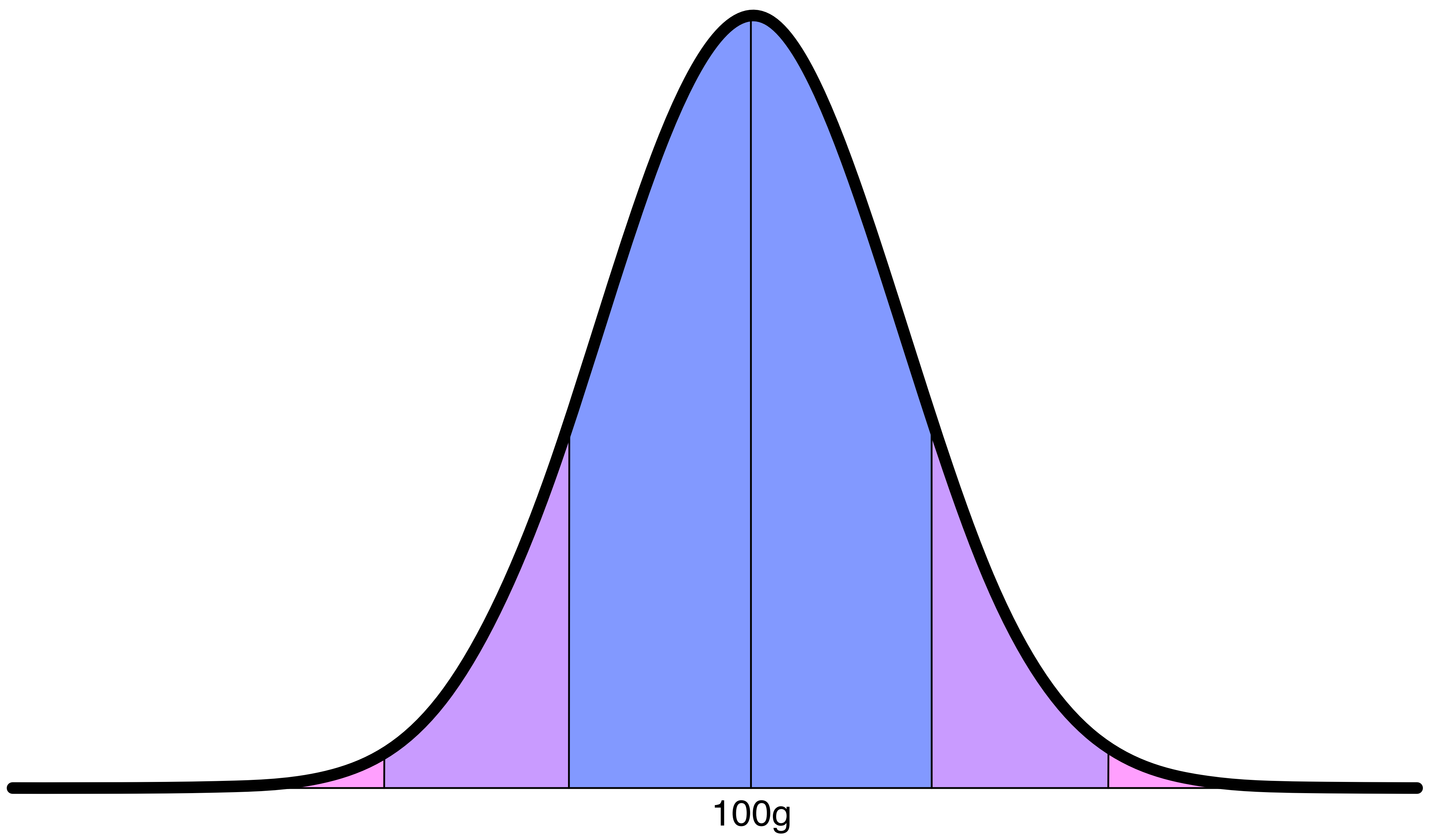


- Suppose the null hypothesis is true (new widgets are from the same distribution as the original widgets)
- Then the sampling distribution should have its mean at 100g
- And the sampling distribution should have a standard deviation of:

$$\frac{\sigma}{\sqrt{N}} = \frac{22}{10} = 2.2$$

  - This is called the **standard error** (SE)
  - Remember, σ is from the population, which we sometimes have to estimate with s (and use a different distribution)

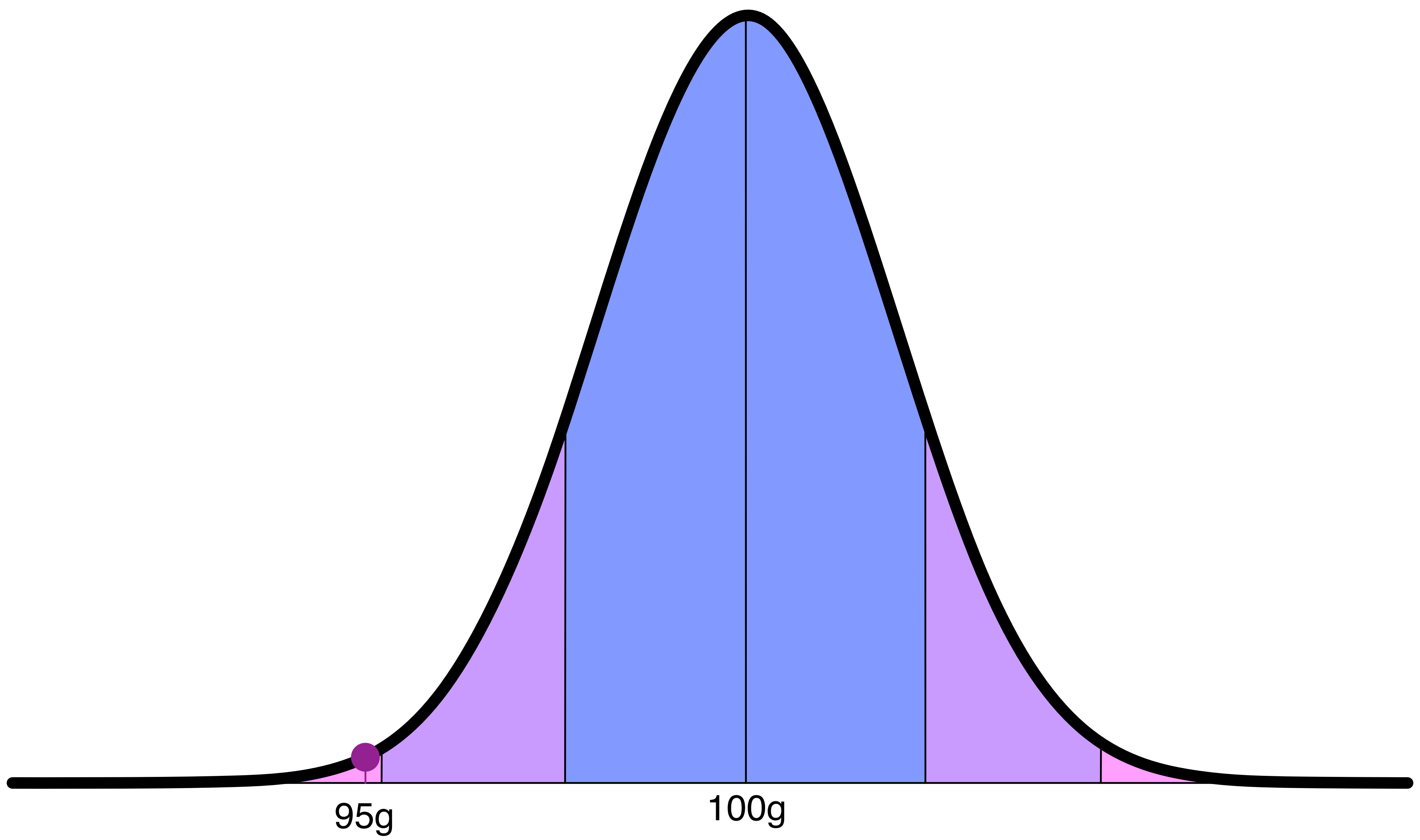
hypothesis testing



- Remember properties of normal distribution:
  - ~68% of points within one σ of μ
  - ~95% of points within two σ of μ
  - ~99.7% of points within three σ of μ

hypothesis testing

- So what about our sample $\bar{x}$ of 95g?



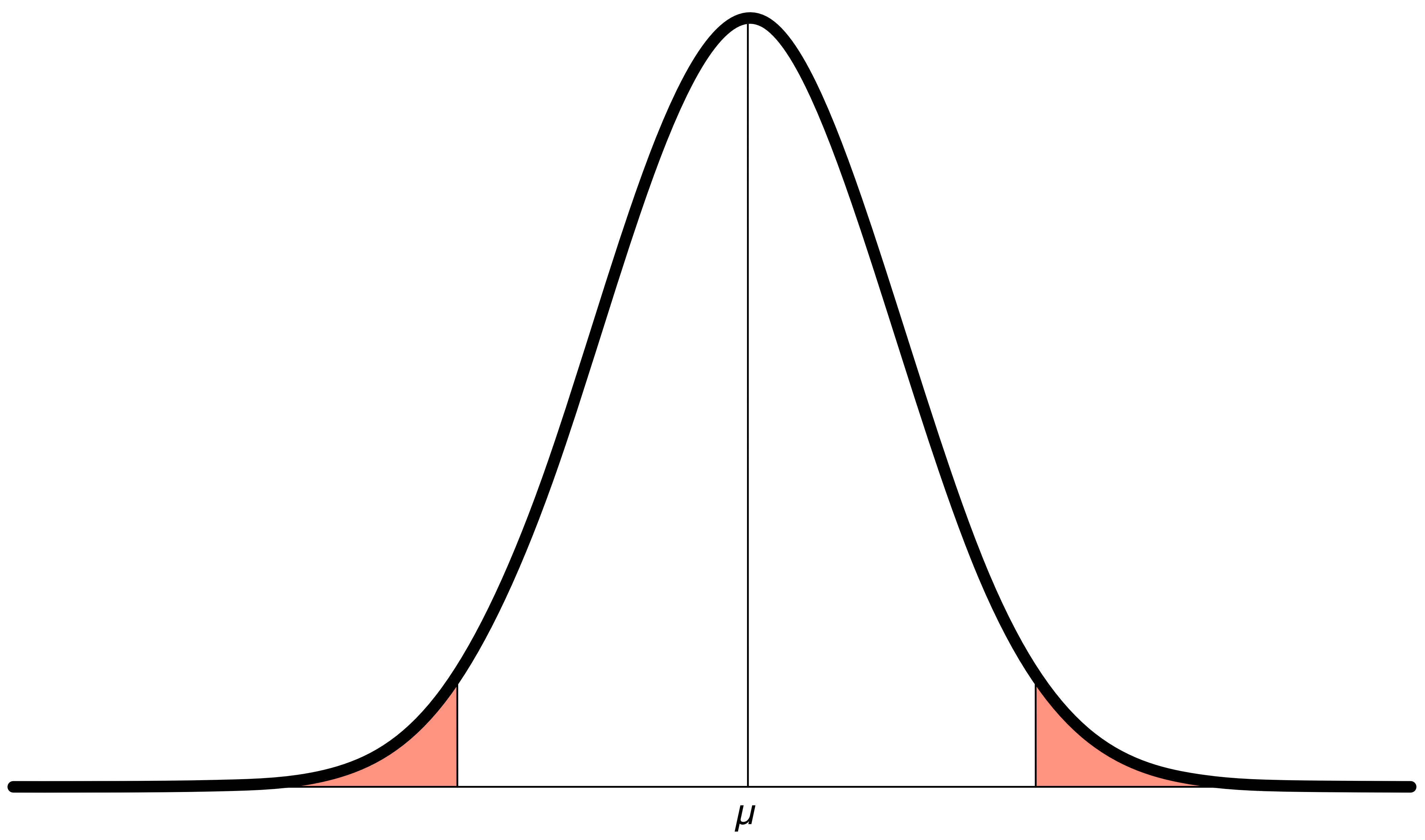
- Very unlikely for it to have come from this distribution!

- Remember properties of normal distribution:
  - ~68% of points within one σ of μ
  - ~95% of points within two σ of μ
  - ~99.7% of points within three σ of μ

z-test

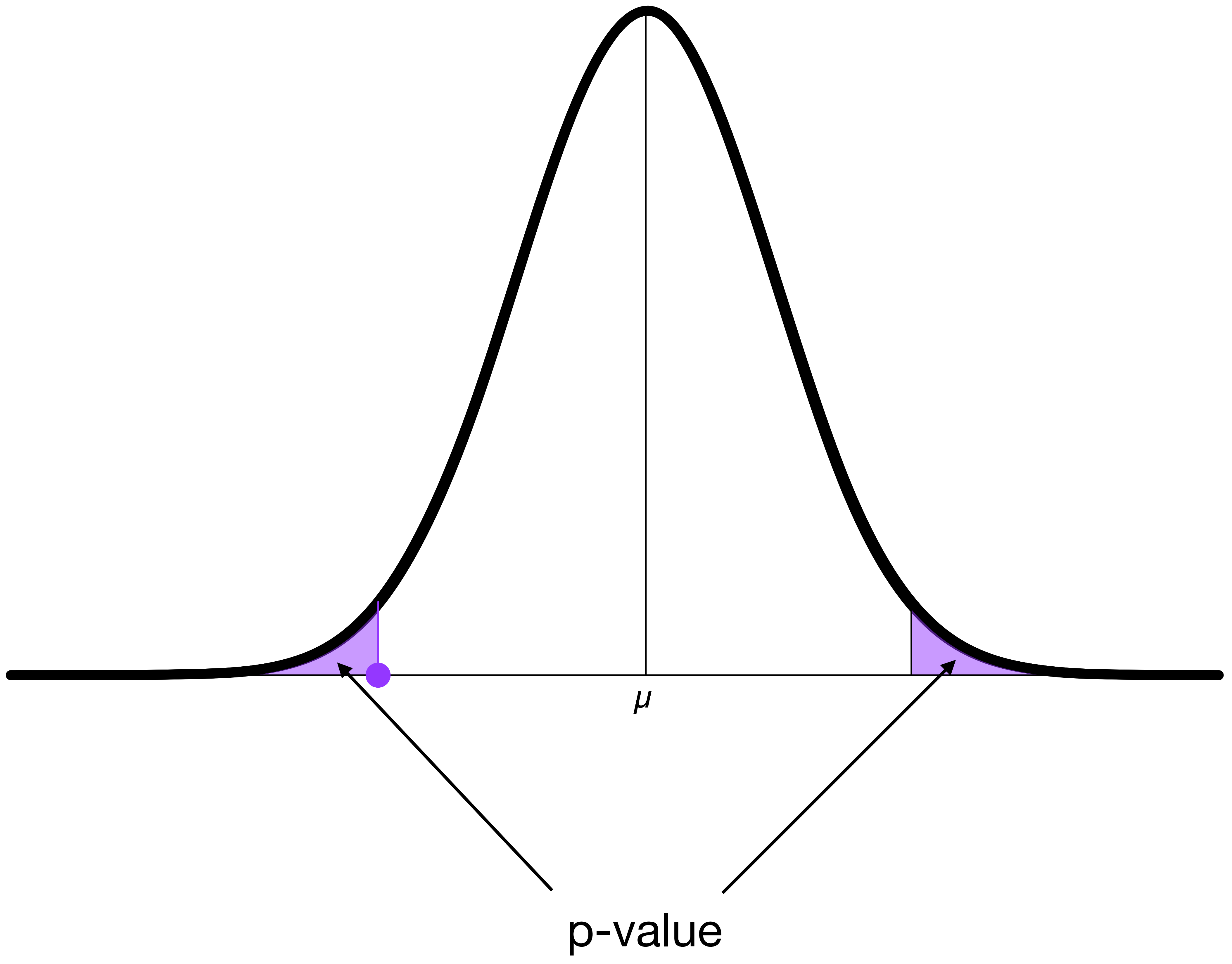
- Common way of dealing with these types of hypotheses is the **z-test**
  - Reasoning about μ
  - When we know σ or if n is large enough (if we don't know σ and n is large enough, we can estimate with *s*)
  - Can construct sampling distribution assuming null hypothesis is true
- Set a **significance level** α for the test
  - Fraction of distribution in "tails" considered anomalous is 2α
  - See whether sample $\bar{x}$ falls in that tail
  - If so, **reject** null hypothesis H_0 in favor of alternative H_1; otherwise, **do not reject** (but this does not prove that H_0 is true)

z-test



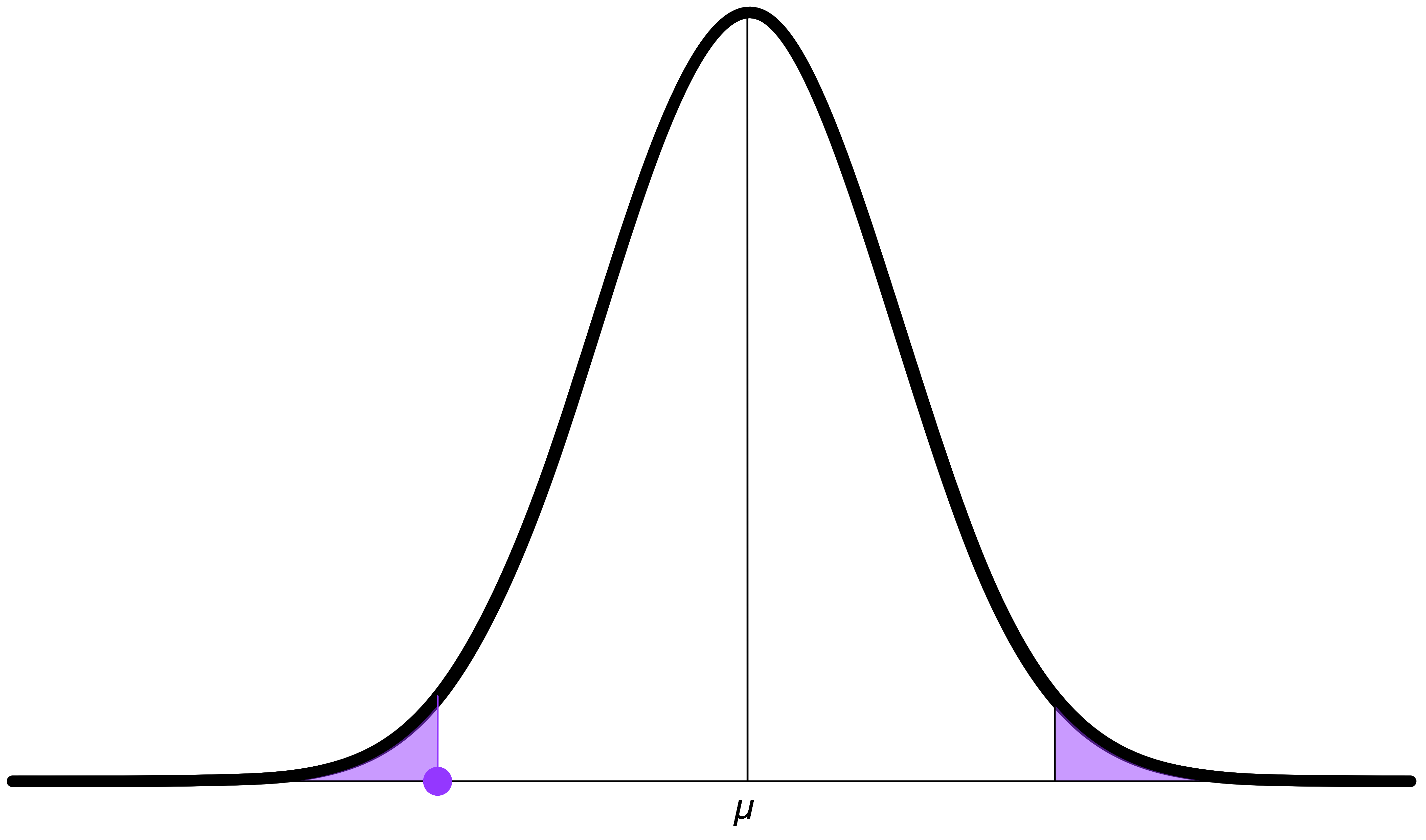
- Set a **significance level** α for the test
  - Fraction of distribution in “tails” considered anomalous is 2α
  - See whether sample $\bar{x}$ falls in that tail
  - If so, **reject** null hypothesis H_0 in favor of alternative H_1; otherwise, **do not reject** (but this does not prove that H_0 is true)

p-value for z-test



- Slightly different way of thinking about the problem
- Place sample $\bar{x}$ on distribution
- Ask what fraction of distribution is farther from the mean μ than the sample $\bar{x}$
- This is your **p-value**
    - Usually ask for p-value < 0.05 or 0.01
    - Sometimes p-value < 0.1 is OK

p-value for z-test



- Procedure:
  - Compute sample mean $\bar{x}$
  - Compute standard deviation of sampling distribution (standard error)

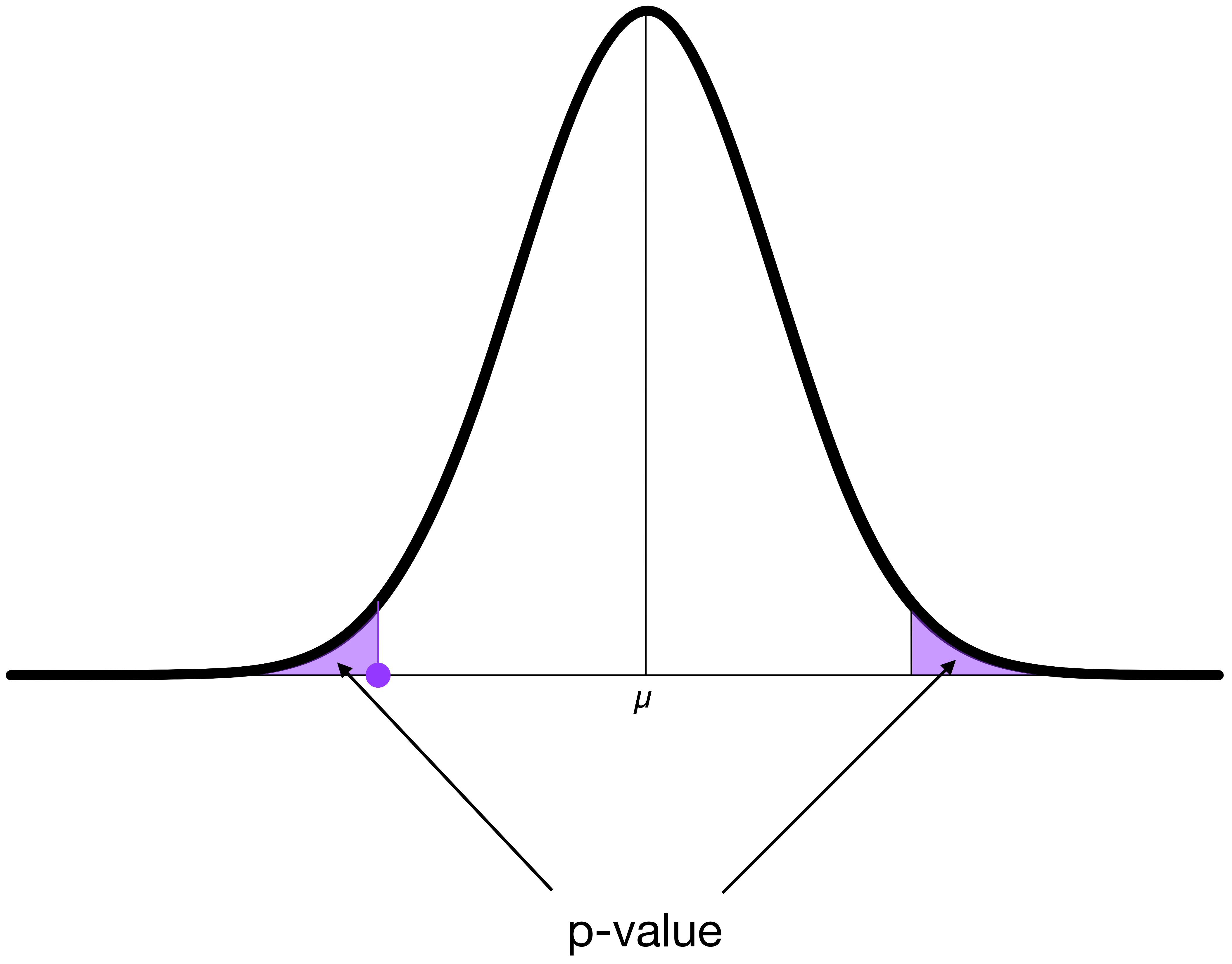
$$SE = \frac{\sigma}{\sqrt{n}}$$

  - Compute **z-score**

$$z = \frac{\bar{x} - \mu}{SE}$$

    - Normalizing the sample to the **standard normal distribution** $\mathcal{N}(0,1)$
  - Compute p-value from z-score

computing p-value from z-score



- One way: look up in a standard table
- In Python:

```
import scipy as sp

# compute z = (x − mu) / SE

p = 2 * sp.stats.norm.cdf(z)
```

comparing two means

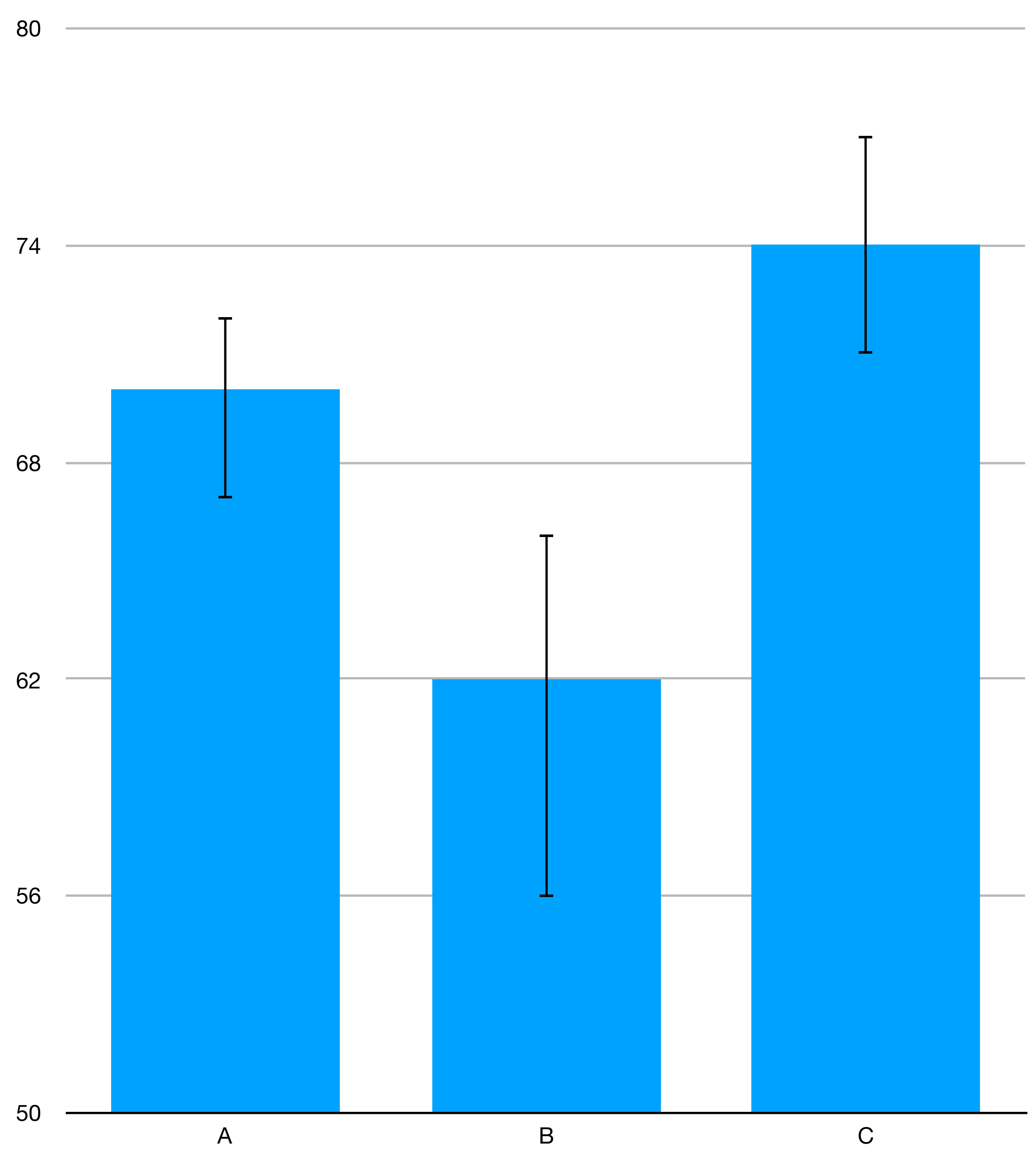
- What if you have *two* sample means and you want to know if their difference is statistically significant?
  - Sample 1: Sample size n_0, mean μ_0, variance σ_0
  - Sample 2: Sample size n_1, mean μ_1, variance σ_1
- Hypotheses
  - H_0 : The means are the same, i.e., $\mu_0 = \mu_1$
  - H_1 : The means are different, i.e., $\mu_0 \neq \mu_1$
- Can use **two-sample z-test**
- Sampling distribution of *difference between two means* has:

$$\mu = \mu_0 - \mu_1 \qquad \sigma = \sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}}$$

- Test point is $\bar{x} = \bar{x}_0 - \bar{x}_1$
- z-score is $(\bar{x} - \mu)/\sigma$

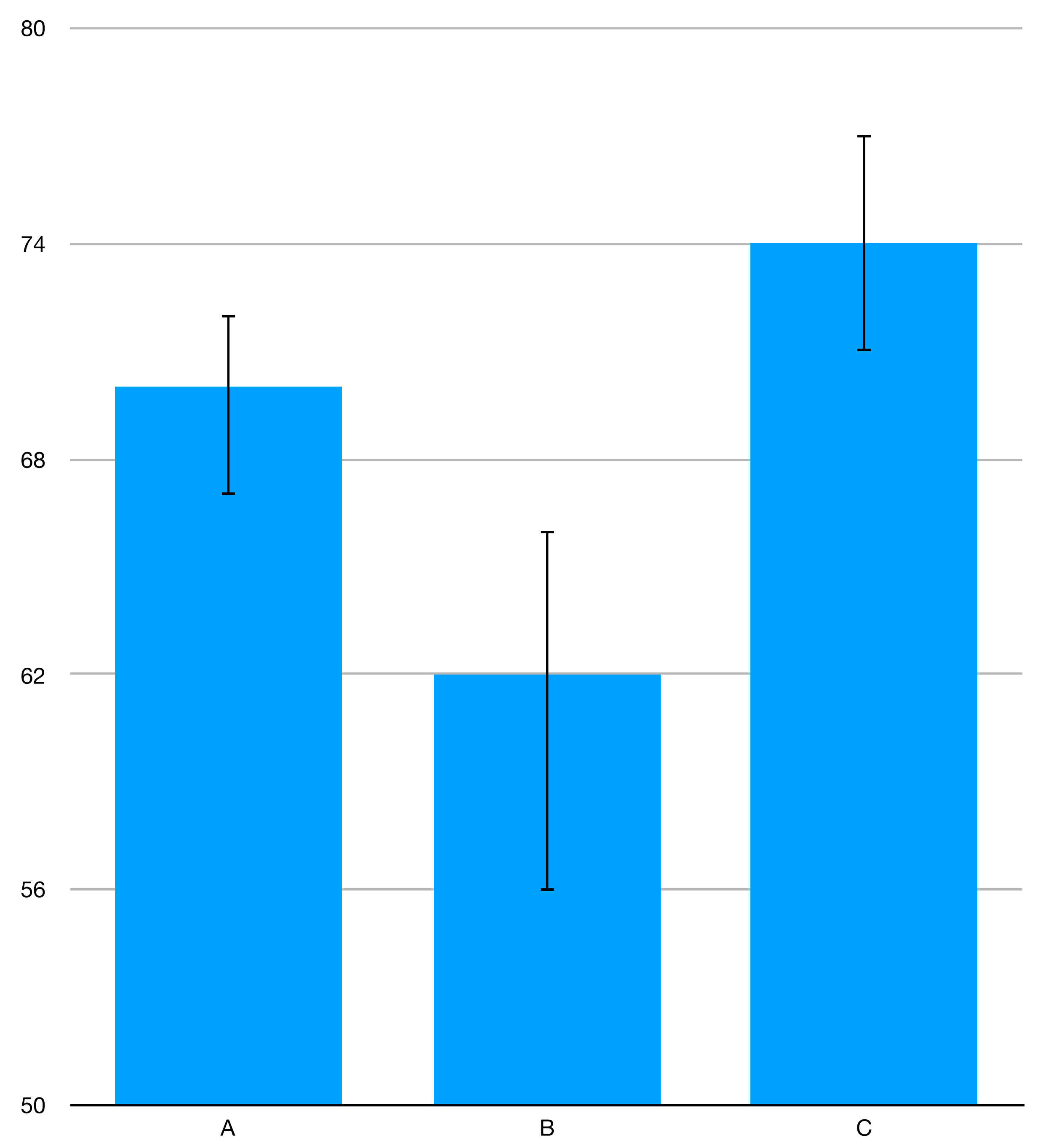
confidence intervals

- We see these a lot: ranges above and below values
  - What do they mean?
- Surprisingly tricky question to answer



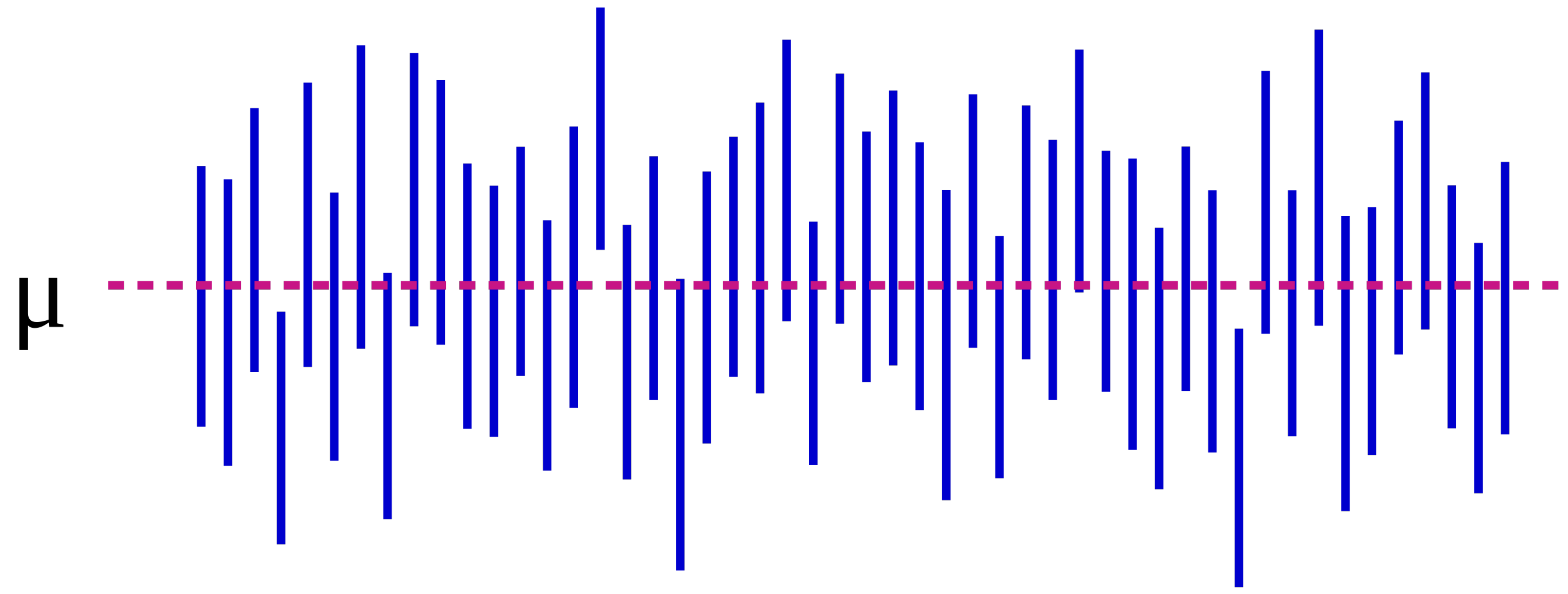
confidence intervals

- A **confidence interval** is a range around the mean which says something about how "good" your estimation procedure is
  - How "good" is your choice of number of samples, given the variance in the population
- Interpretation of a confidence interval:
  - **if I were to repeat the experiment a large number of times, 95 percent of confidence intervals would contain the population mean** or
  - **when I run the experiment, there is a 95 percent chance that the population mean will fall within the confidence interval** or
  - **if the population mean is inside the confidence interval, it would not be statistically significant (informally, you wouldn't be surprised!)**



confidence intervals

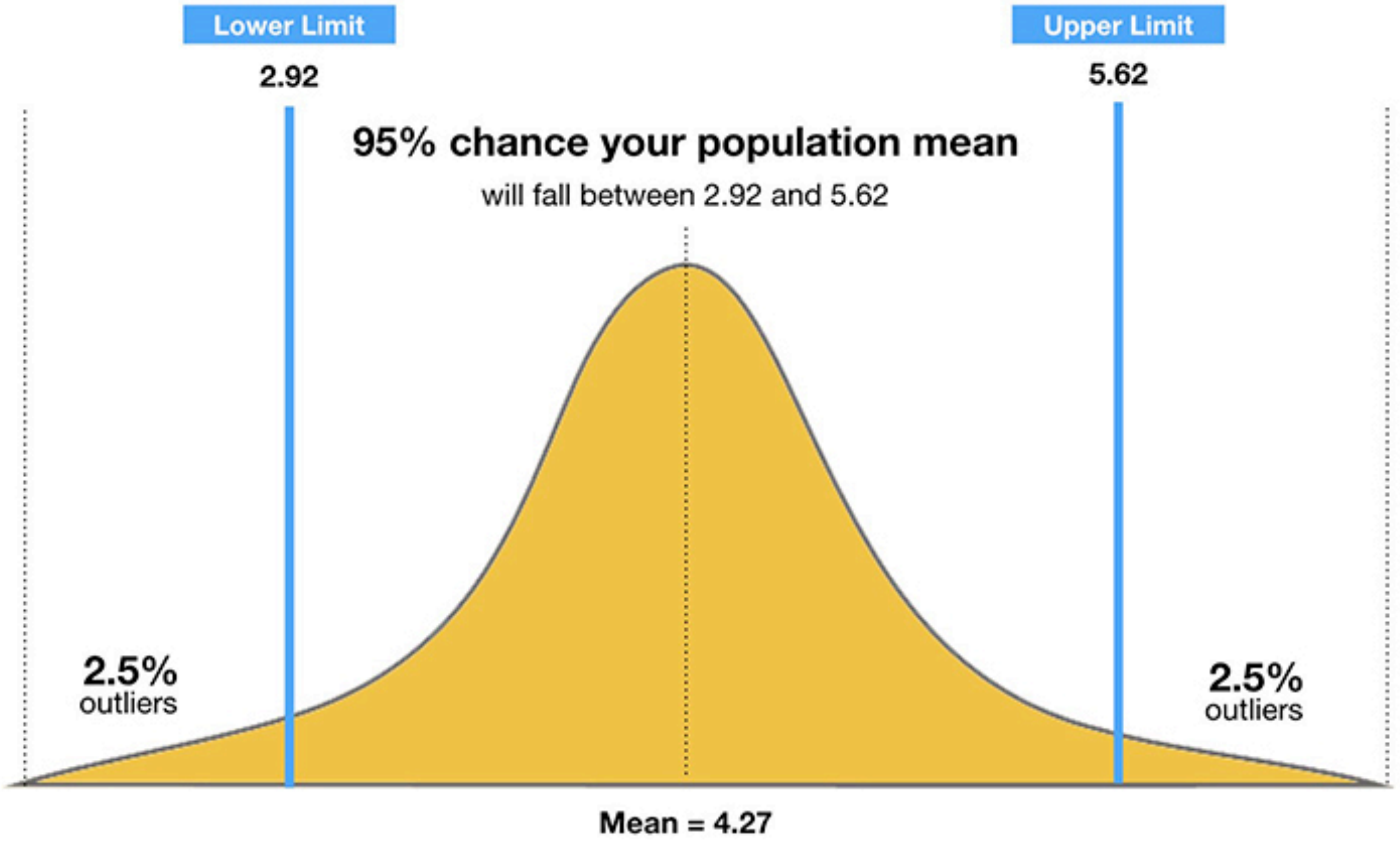
- A **confidence interval** is a range around the mean which says something about how “good” your estimation procedure is
    - How “good” is your choice of number of samples, given the variance in the population
- Interpretation of a confidence interval:
    - **if I were to repeat the experiment a large number of times, 95 percent of confidence intervals would contain the population mean** or
    - **when I run the experiment, there is a 95 percent chance that the population mean will fall within the confidence interval** or
    - **if the population mean is inside the confidence interval, it would not be statistically significant (informally, you wouldn’t be surprised!)**



source: NYW-confidence-interval.svg
Wikipedia user Tsyplakov

confidence intervals

- If the population parameter is outside the **c**% confidence interval, then an event occurred that had a probability of less than (100 - **c**)% of happening
- Note that we are setting **c** ahead of time (unlike with hypothesis testing, where we figure out how likely/unlikely something is *after* the fact)
    - Wide confidence interval: The variance of your data is high (and/or your sample size is small), so we need a wide interval to make the above statement true.
    - Narrow confidence interval: The variance of your data is small (and/or your sample size is large), so we *don't* need a wide interval to make the above statement true.



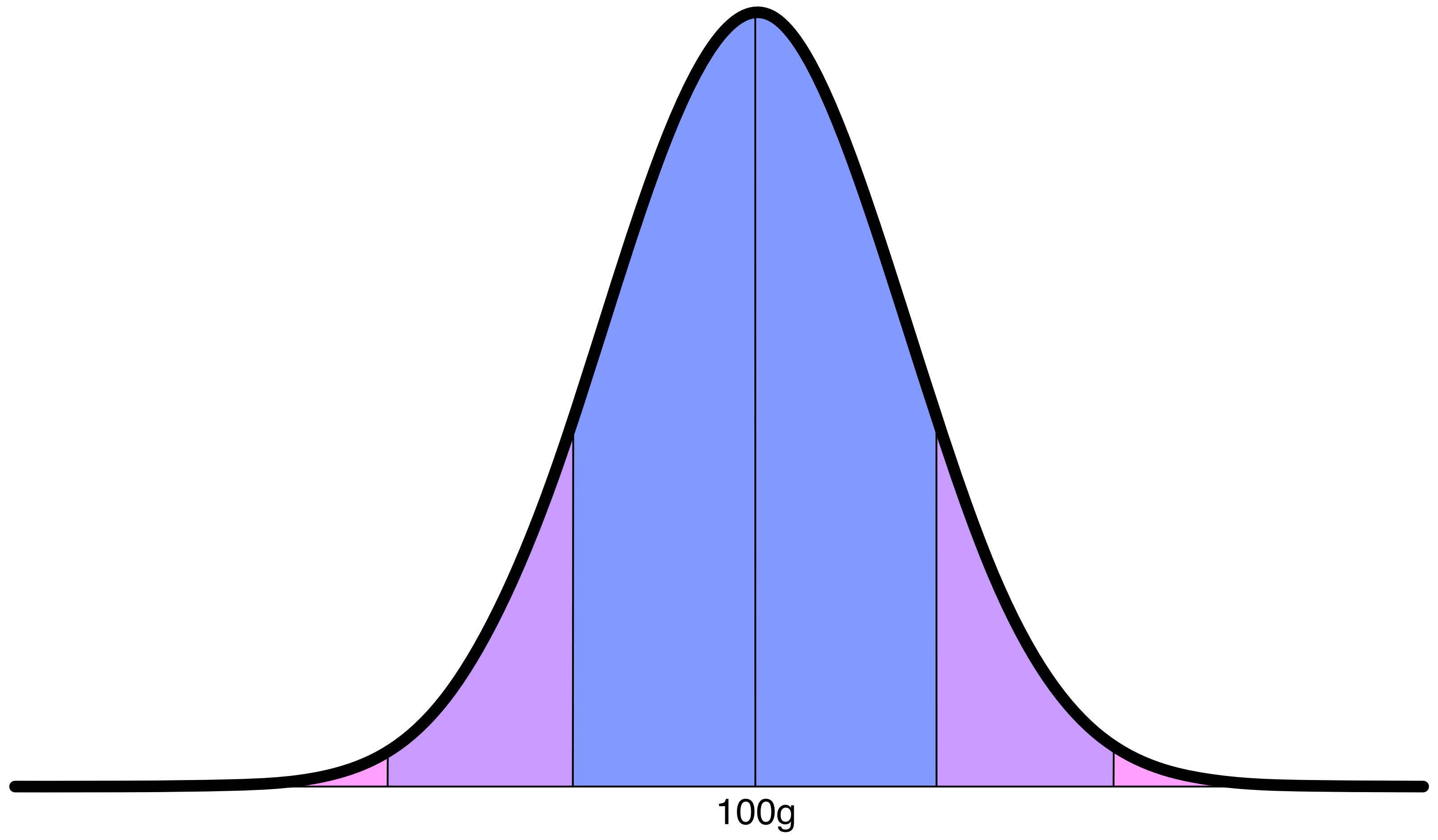
computing confidence intervals

- Conceptually similar to z-tests, except now the sampling distribution is centered around the *sample* mean (instead of the hypothesized population mean

- Remember definition of z-score:

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}}$$

- And p-value:
```
p = 2 * sp.stats.norm.cdf(z)
```



computing confidence intervals

- Conceptually very similar to z-tests, except now *sampling distribution is centered around the sample mean* (instead of the hypothesized population mean)

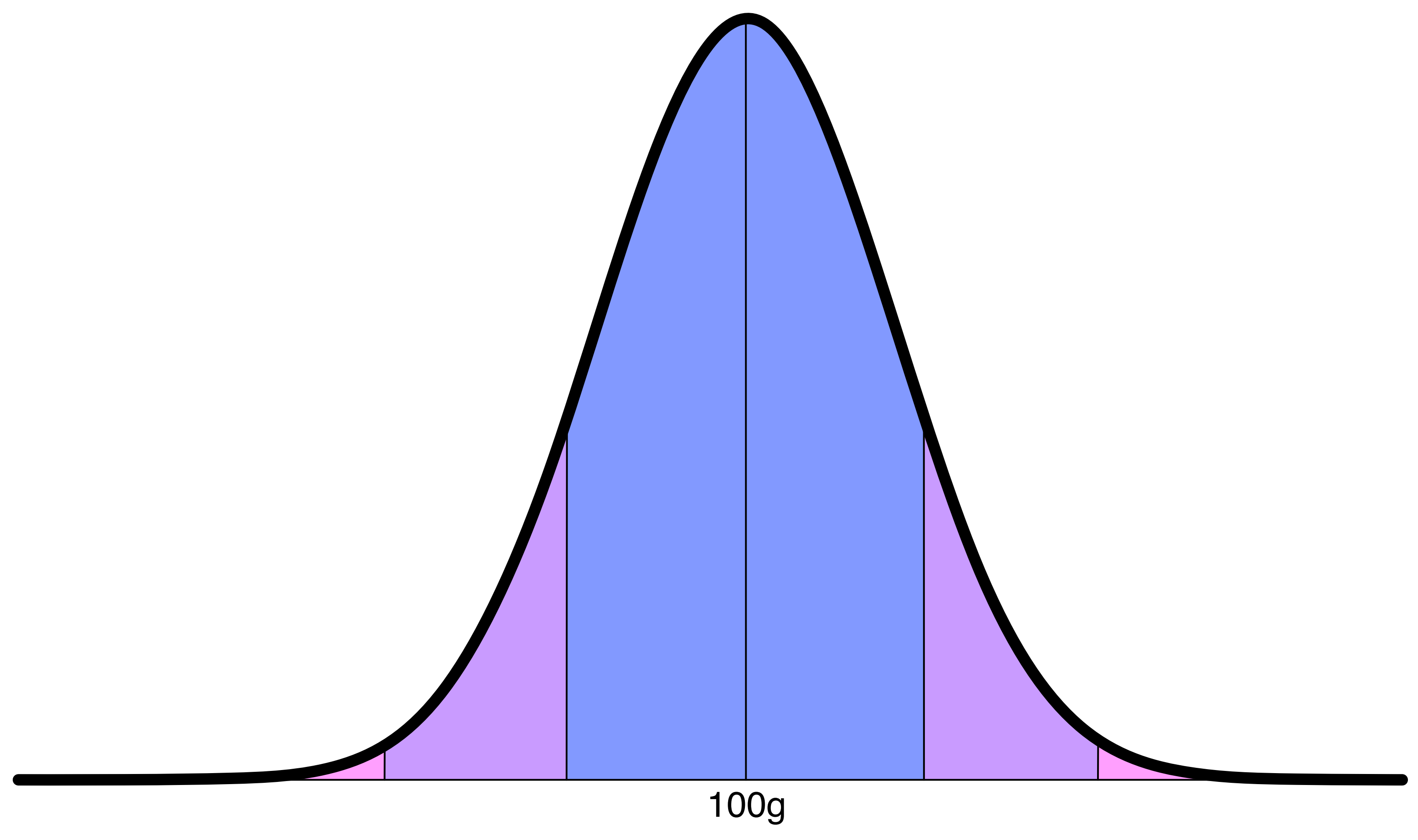
- Remember definition of z-score:

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{N}}$$

- And p-value:

```
p = 2 * sp.stats.norm.cdf(z)
```

- If c is the desired confidence level, what z do we need so that $p \leq (1 - c)$?



what z do we need?

- Call this z_c
- Compute like so:

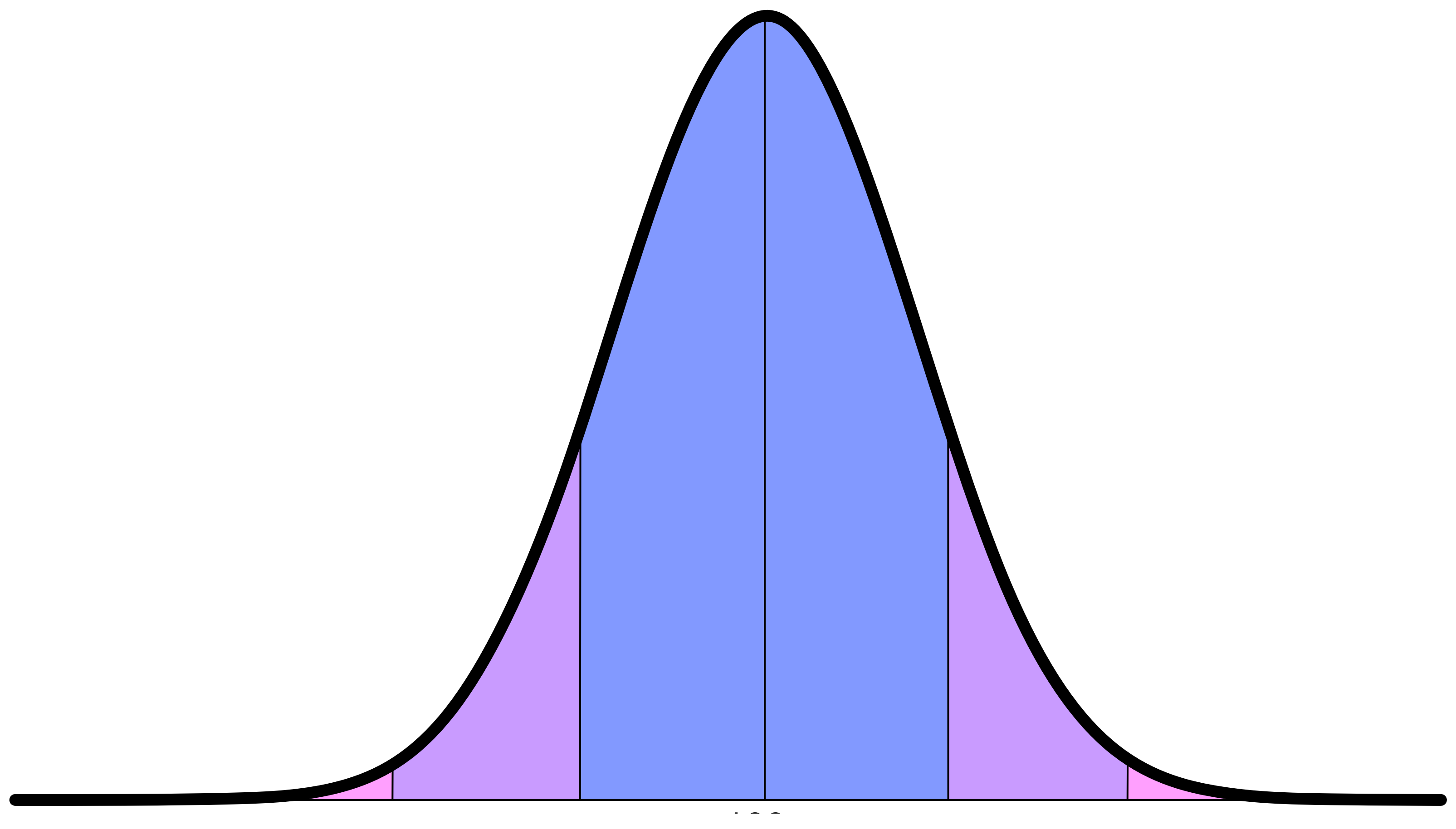
```
z_c = sp.stats.norm.ppf(1 - (1 - c)/2)
```

"Inverse" of the CDF function

- Now we can answer the question: *What range of μ would be "unsurprising" at c% confidence level?*

$$z_c = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}} \rightarrow \mu = \bar{X} \pm \frac{z_c \cdot \sigma}{\sqrt{N}}$$

- This is your c% confidence interval



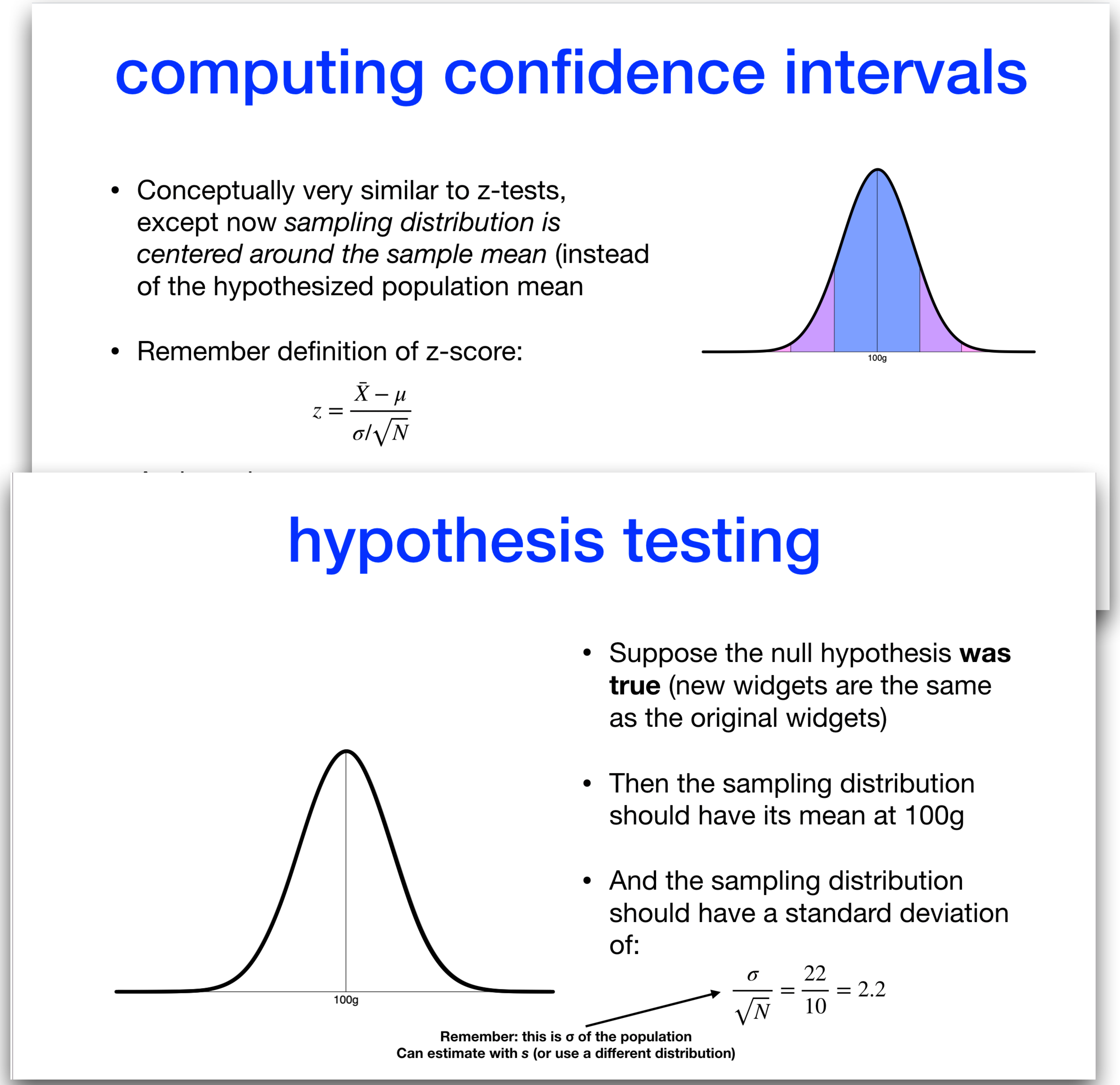
Back to our original example …

$$\bar{x} = 95g, \sigma = 22g, n = 100$$

$90\,\% : (92.42,\ 97.58)$
$95\,\% : (91.38,\ 98.62)$
$99\,\% : (90.69,\ 99.31)$

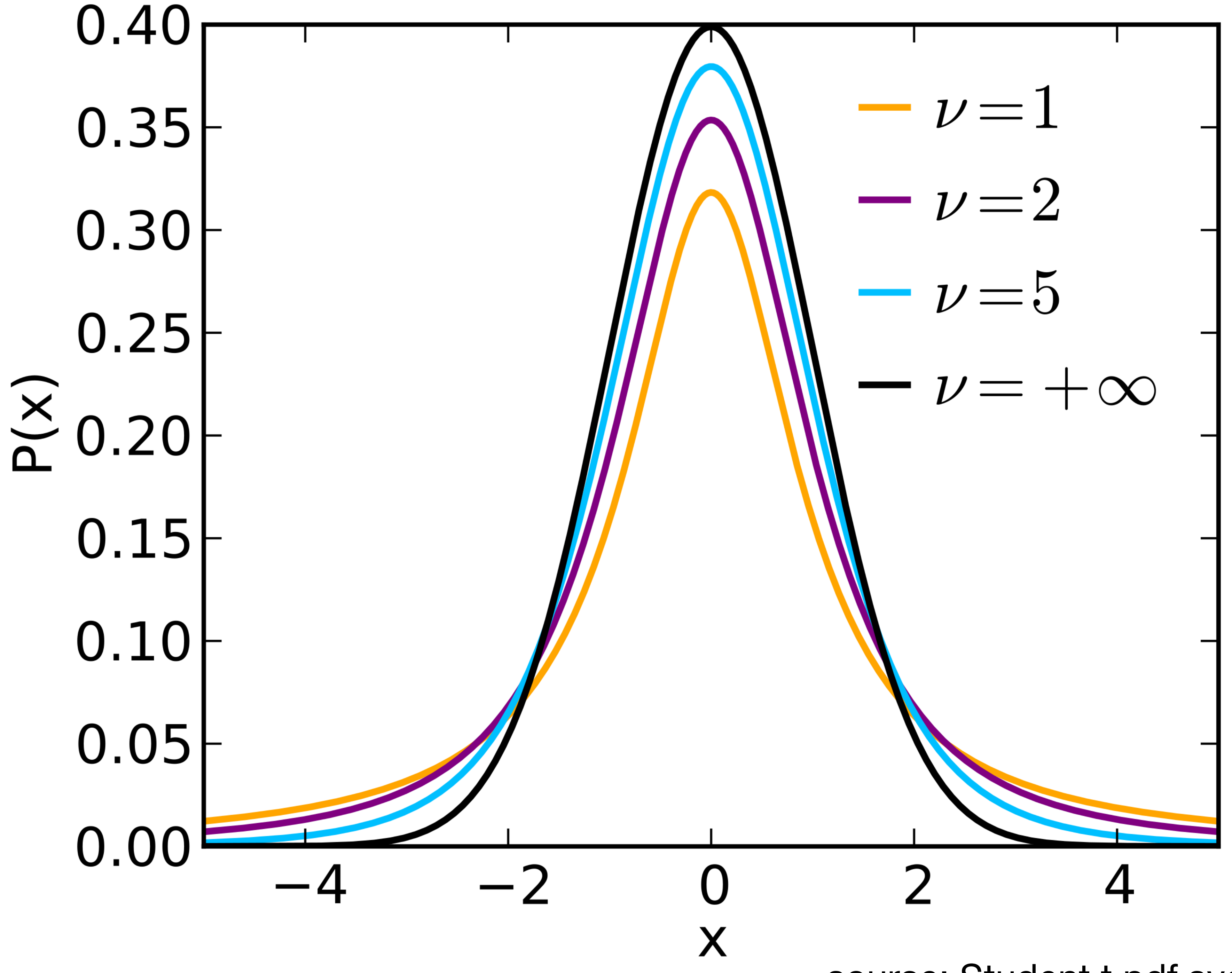
we’ve been fudging

- Recall that to use the z-distribution, we must either know σ or have large enough n
- The **student’s t-distribution** and **t-test** is used when the normal approximation does not hold
  - Also used to reason about μ, including building confidence intervals
  - When we don’t know σ and when $n < 30$



student's t distribution

- Similar to the standard normal distribution
  - Symmetric about mean
  - Bell curve shaped
- But has **fatter tails**, i.e., more weight of the distribution away from the mean
  - Accounts for outliers better
- Parameter on the distribution is the **degrees of freedom** v
  - $v = n - 1$: One less than the number of samples
  - Looks more and more like the standard normal as $n \to \infty$



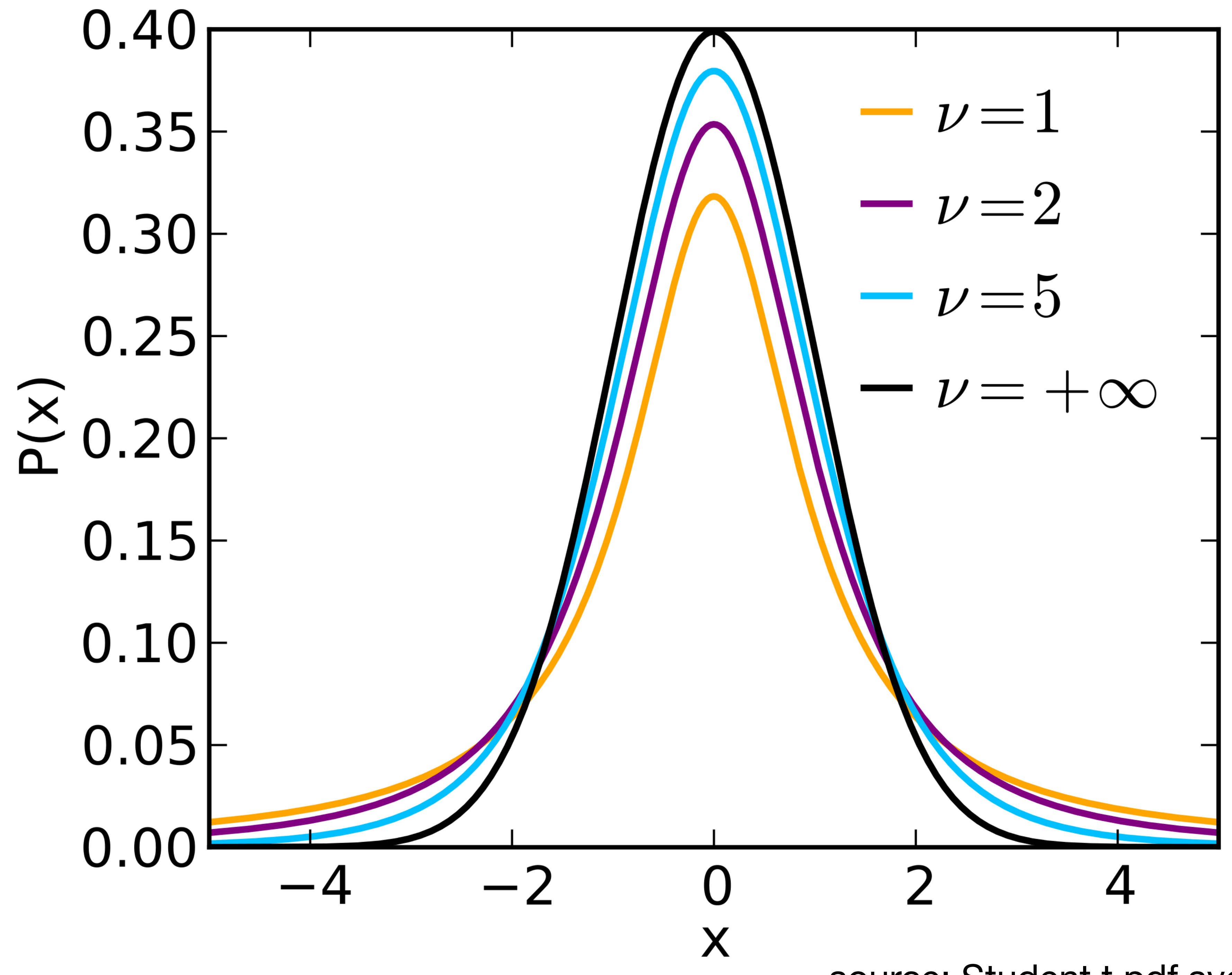
t-test

- Works the same as the z-test, except
  - use s instead of σ
  - compare to the t-distribution
- Computing the test statistic:

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{N}}$$

Compare to the formula for z



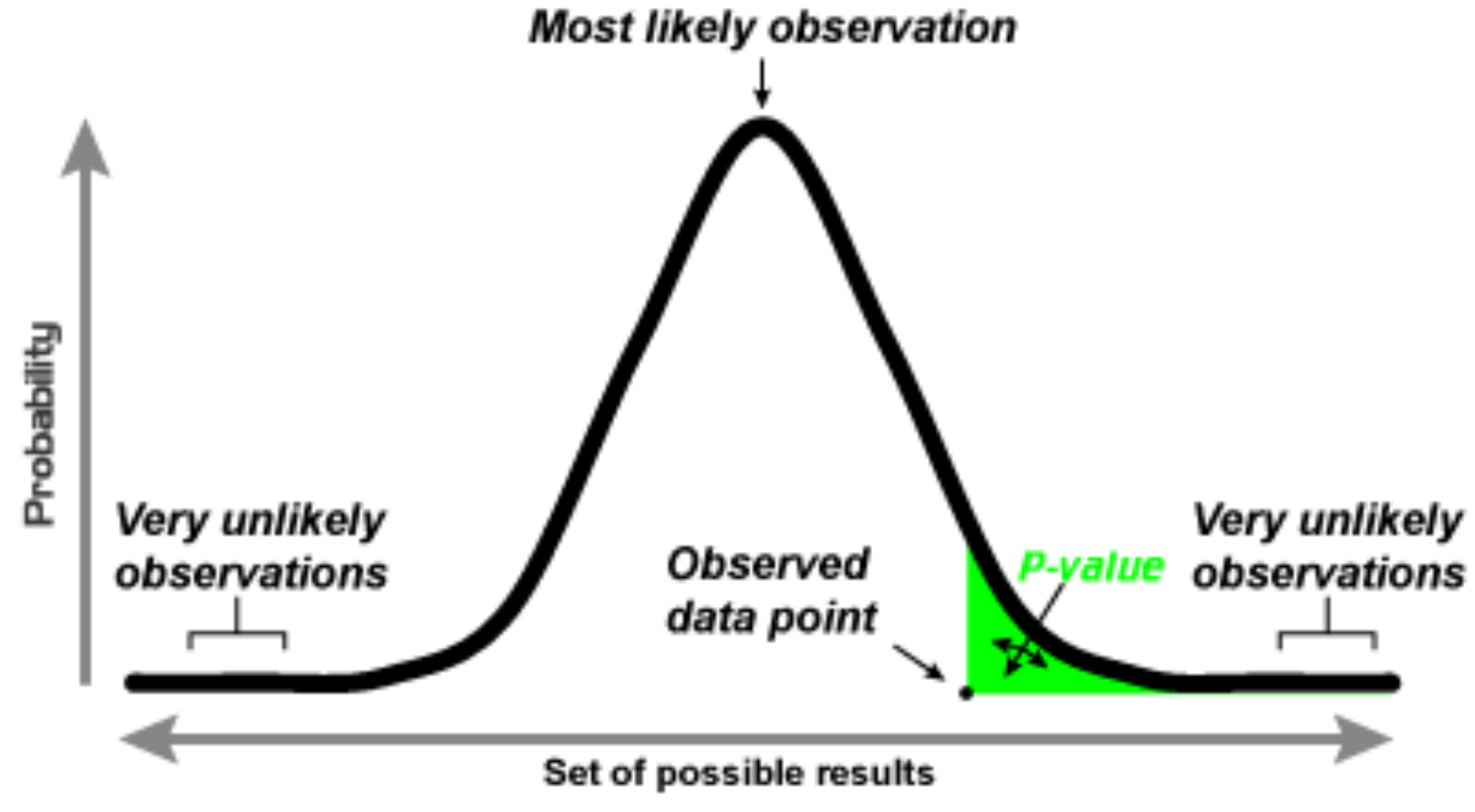
source: Student t pdf.svg
Wikipedia user Skbkekas. Creative Commons 3.0 License

```
p = 2 * sp.stats.t.cdf(t, df)

t_c = sp.stats.t.ppf(1 - (1 - C)/2, df)
```

one-sided tests

- Sometimes we are only interested in values departing from the mean in one direction
    - This is a **one-sided** or **one-tailed test**
- For example, suppose we want to assess whether our widgets are being produced at a significantly *higher* weight:
    - $H_0 : \mu \leq 100g$ ← **Null hypothesis is always the logical "opposite"**
    - $H_1 : \mu > 100g$
- How does the p-value compare between one and two-sided tests?



A p-value (shaded green area) is the probability of an observed (or more extreme) result arising by chance

- Any given datapoint has *half* the p-value in a one-sided test than it does in a two-sided test

simple extensions

- What do we do in a two-sample test when one of the samples violates the normal approximation assumptions?
  - Use a **two-sample t-test**
- Can we build a confidence interval around a mean when the normal approximation is violated?
  - Yes, just use the t-statistic in place of the z-score
- What if we are only interested in a confidence interval on one side (e.g., a lower bound or an upper bound)?
  - Can use a **one-sided interval**, where one of the bounds is replaced by $-\infty$ or $+\infty$