ECE 20875

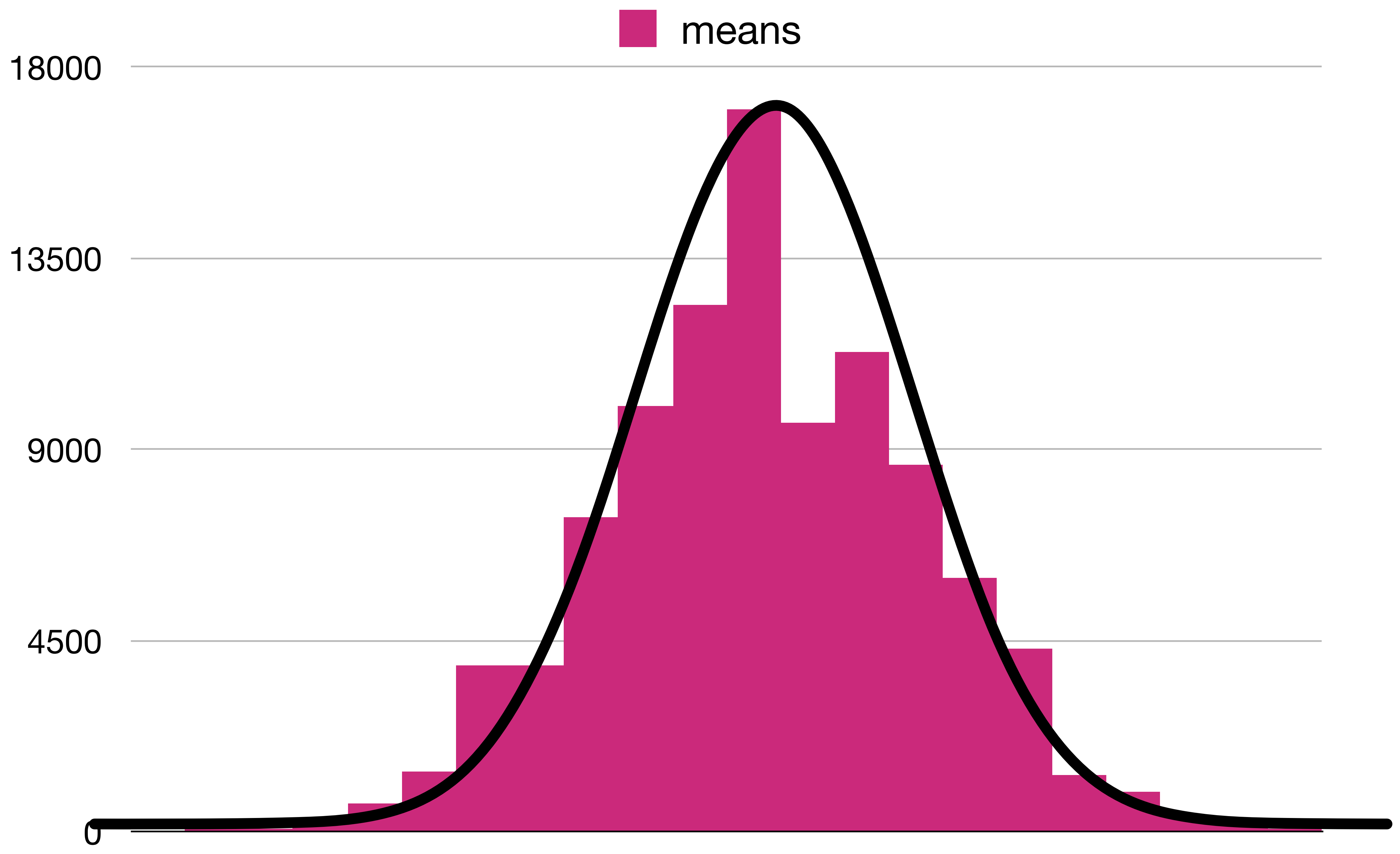
Python for Data Science

David Inouye and Qiang Qiu

(Adapted from material developed by Profs. Milind Kulkarni, Stanley Chan, Chris Brinton, David Inouye, Qiang Qiu)

confidence intervals and hypothesis testing

sampling distribution

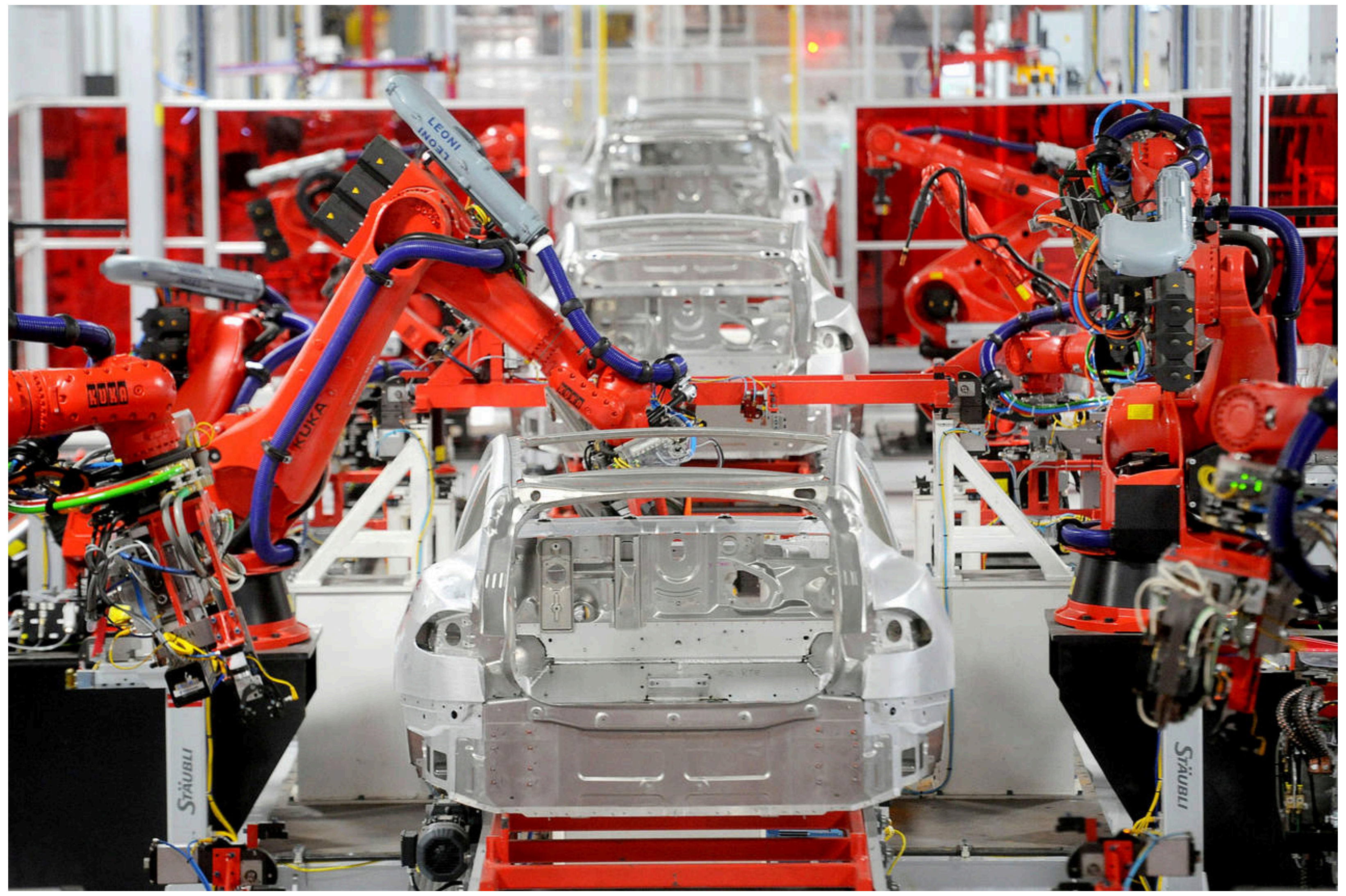


Each data point is the $\bar{x}$ of one experiment

- 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

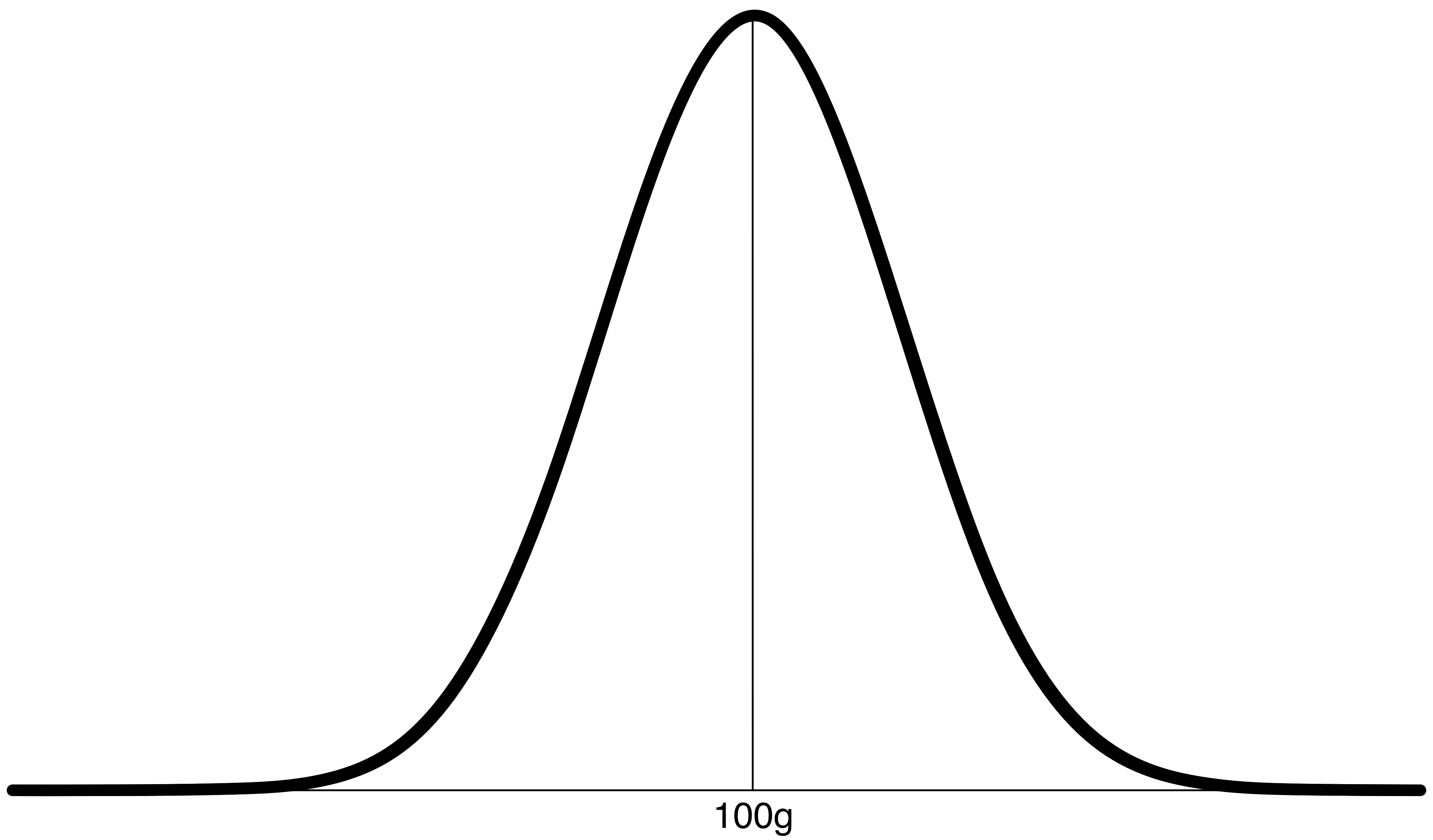
- Suppose a factory claims to produce widgets with an average weight of 100g and a standard deviation of 22g
- We receive a new shipment of widgets which seem off, and we want to see whether the factory has shifted
- Form two hypotheses:
  - **Null hypothesis** (H_0): The factory is producing according 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



$\mathcal{N}(100, 2.2)$

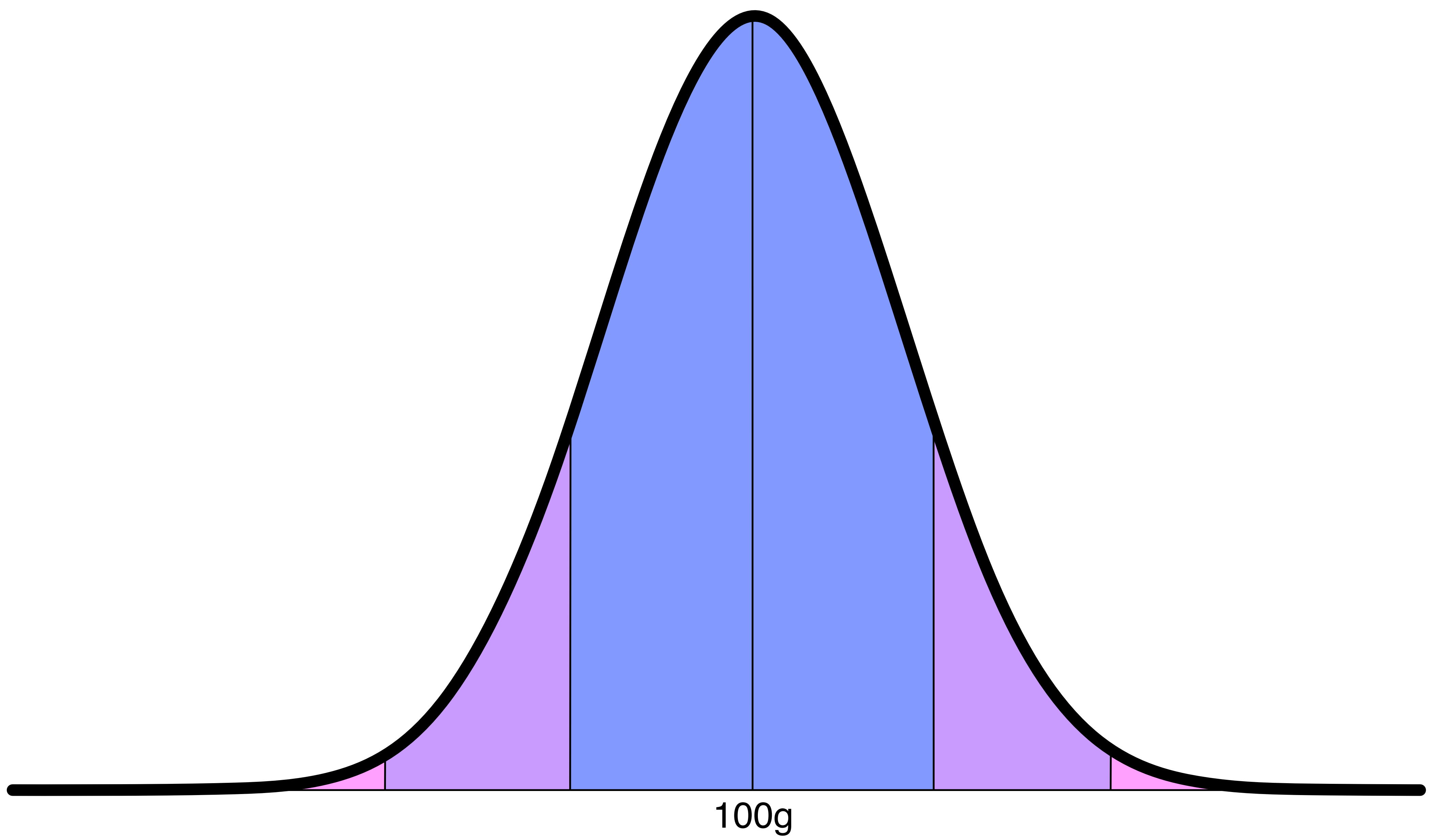
- Suppose the null hypothesis is true (new widgets are from the same distribution as the original widgets)
- Then the sampling distribution should have its mean at $\mu = 100\text{g}$
- And the sampling distribution should have a standard deviation of:

$$SE \triangleq \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \approx \frac{22}{10} = 2.2\text{g}$$

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

hypothesis testing

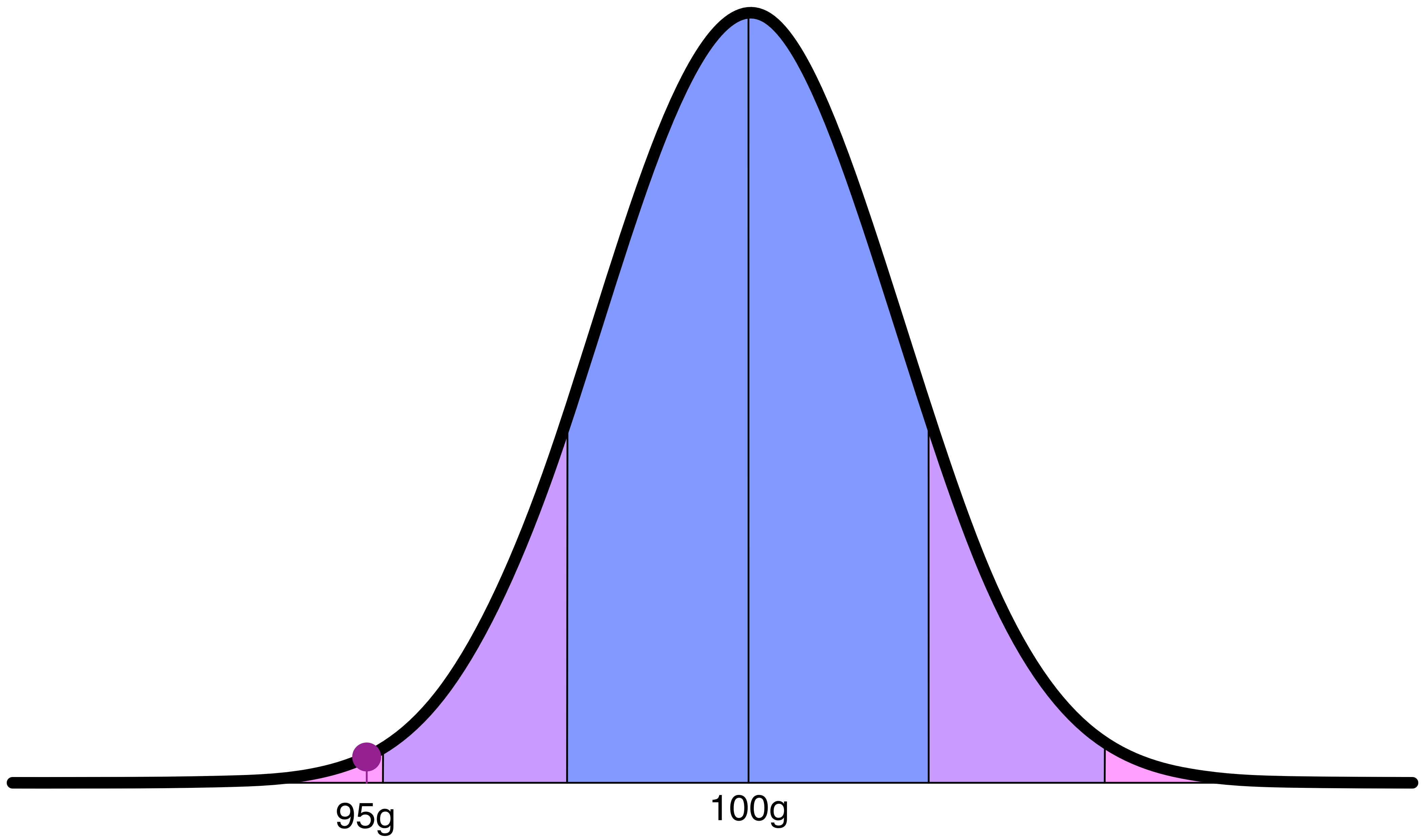
$\mathcal{N}(100, 2.2)$



- 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

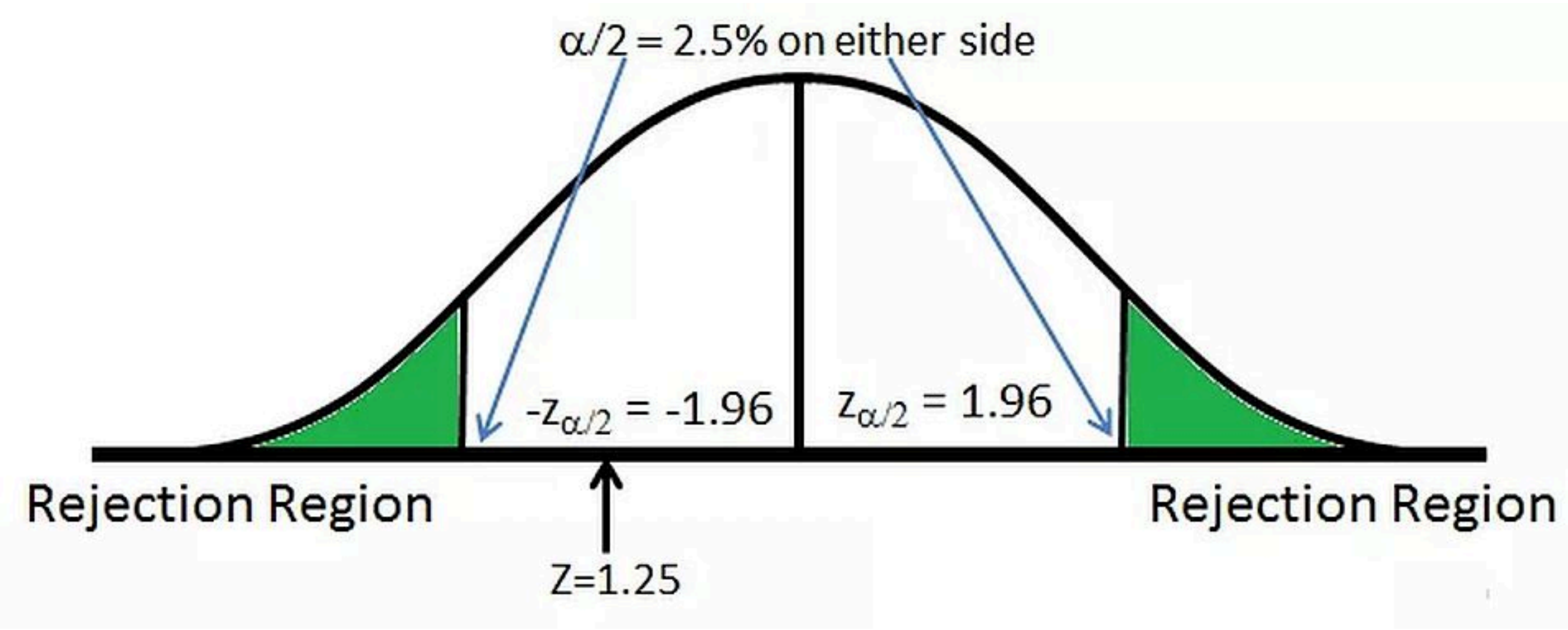
- So what about our sample $\bar{x}$ of 95g?



- 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
- 95g is between 2 and 3 $\sigma_{\bar{X}}$ of μ

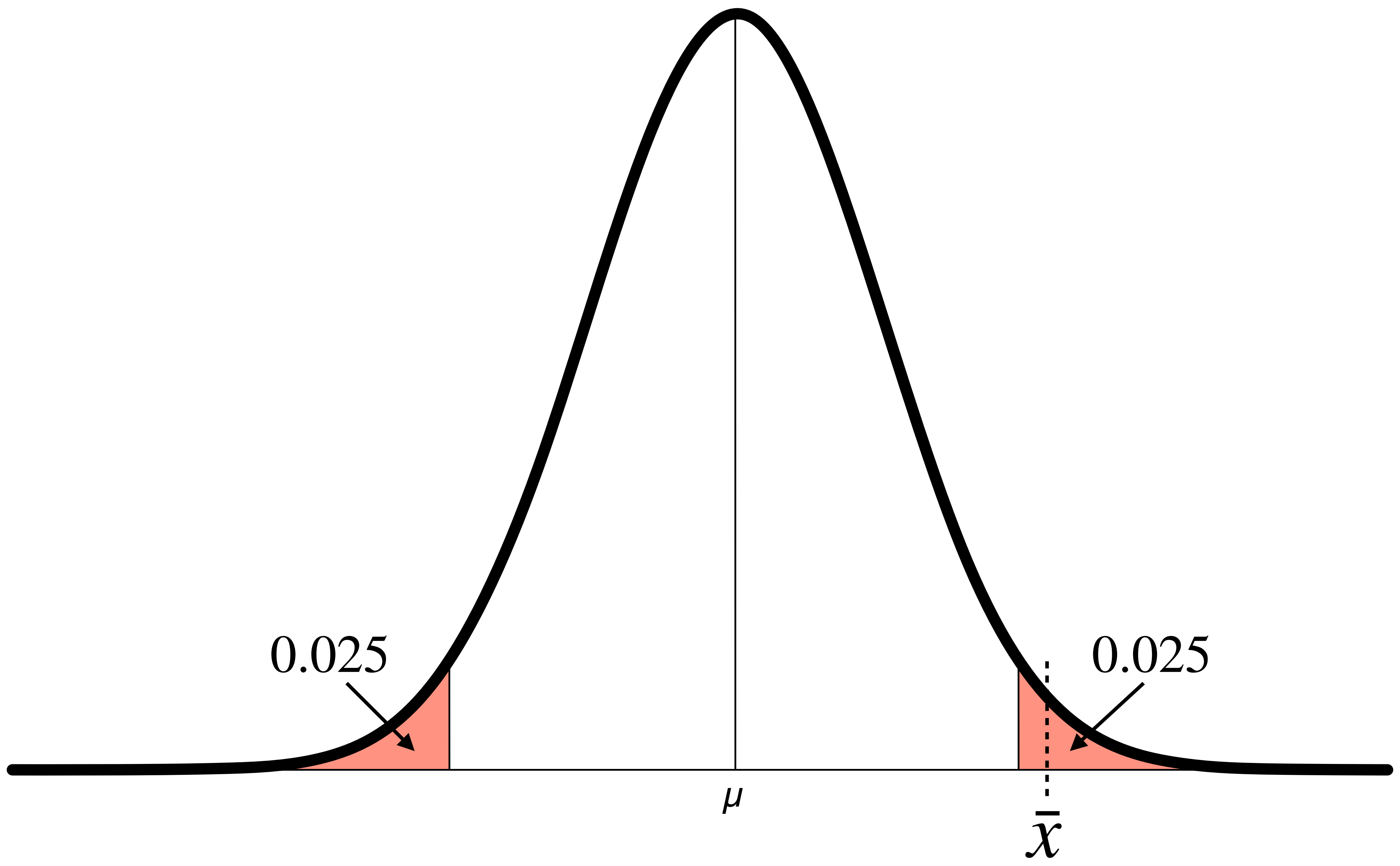
z-test



- The statistical **z-test**
  - Reasoning about μ
  - Applicable 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 each "tail" considered anomalous is $\alpha/2$ (if **two-sided test**)
  - 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

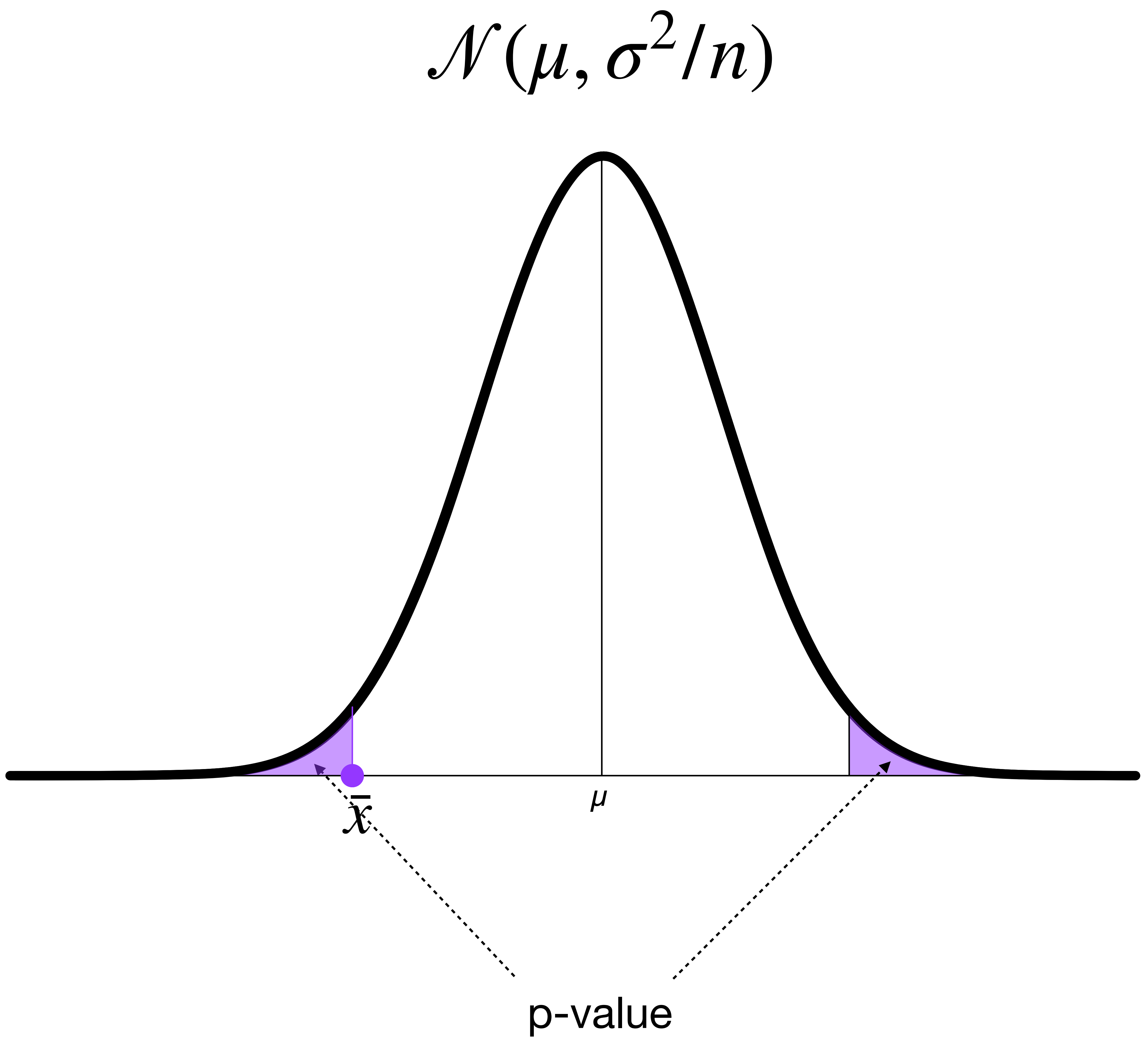
Significance level: $\alpha = 0.05$



In this case, reject H_0

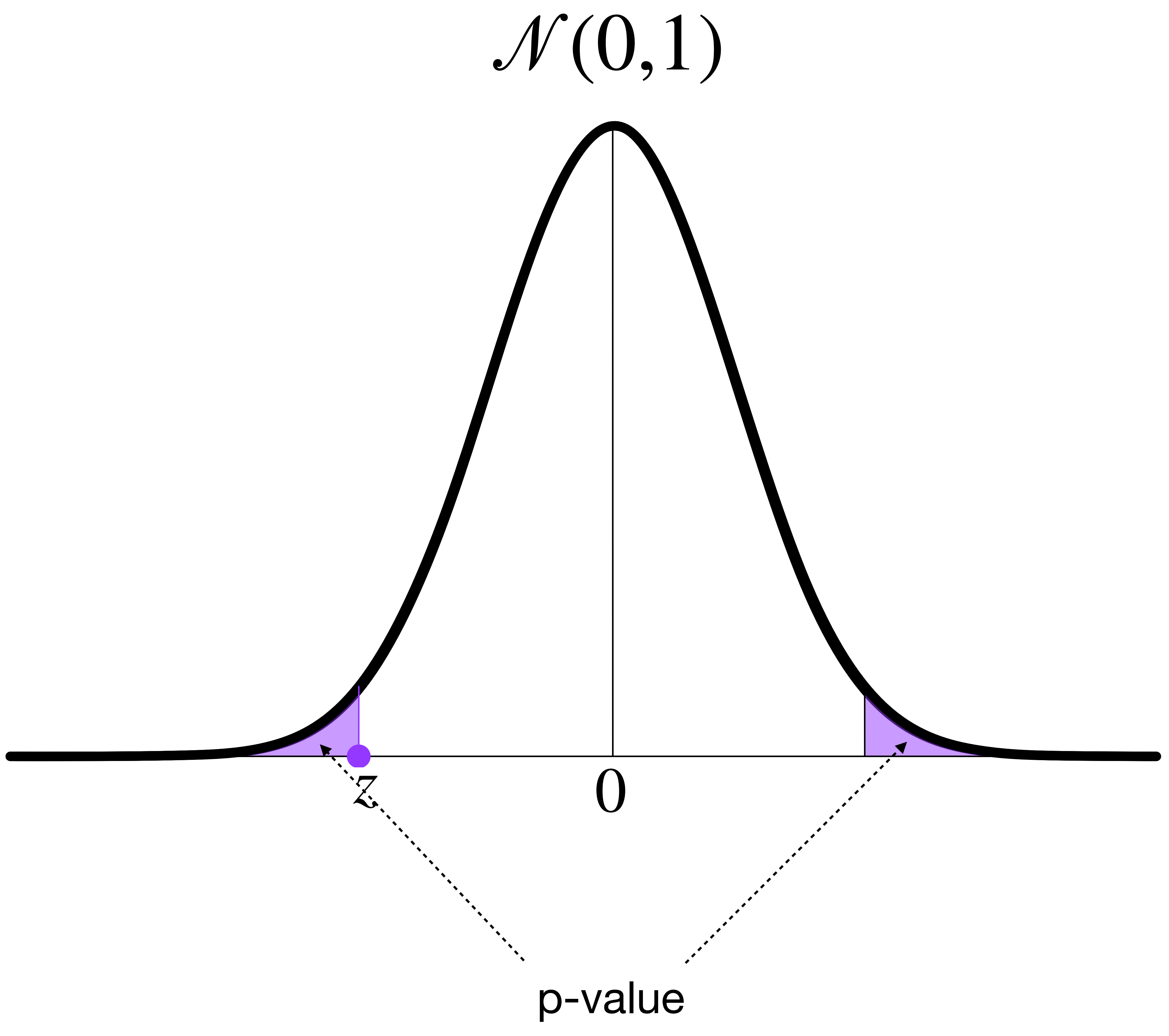
- Set a **significance level** α for the test
  - Fraction of distribution considered anomalous is $\alpha/2$ in each "tail" (if two-sided)
  - See whether sample $\bar{x}$ falls in that tail
  - If so, *reject* null hypothesis H_0 in favor of alternative H_1; otherwise, *do not reject* (but this does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

p-value for z-test



- We can formalize this logic by calculating the **p-value**
- 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, which is compared to the significance level α:
  - Usually ask for $\alpha = 0.05$ or 0.01 (i.e., so that $p \leq 0.05, 0.01$ for significance)
  - Sometimes $\alpha = 0.1$ is OK

procedure



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

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

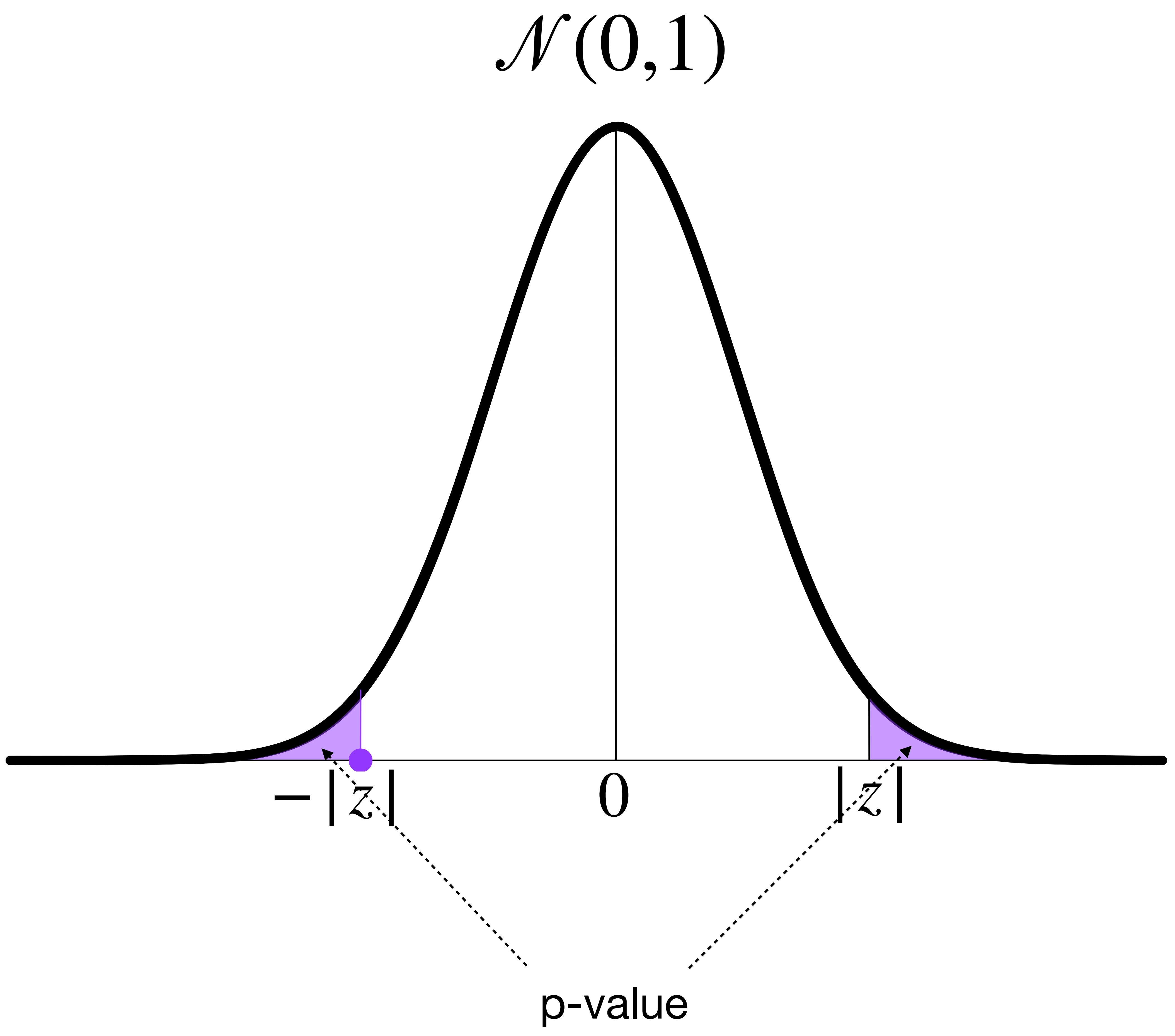
- Compute **z-score**

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

  - 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:

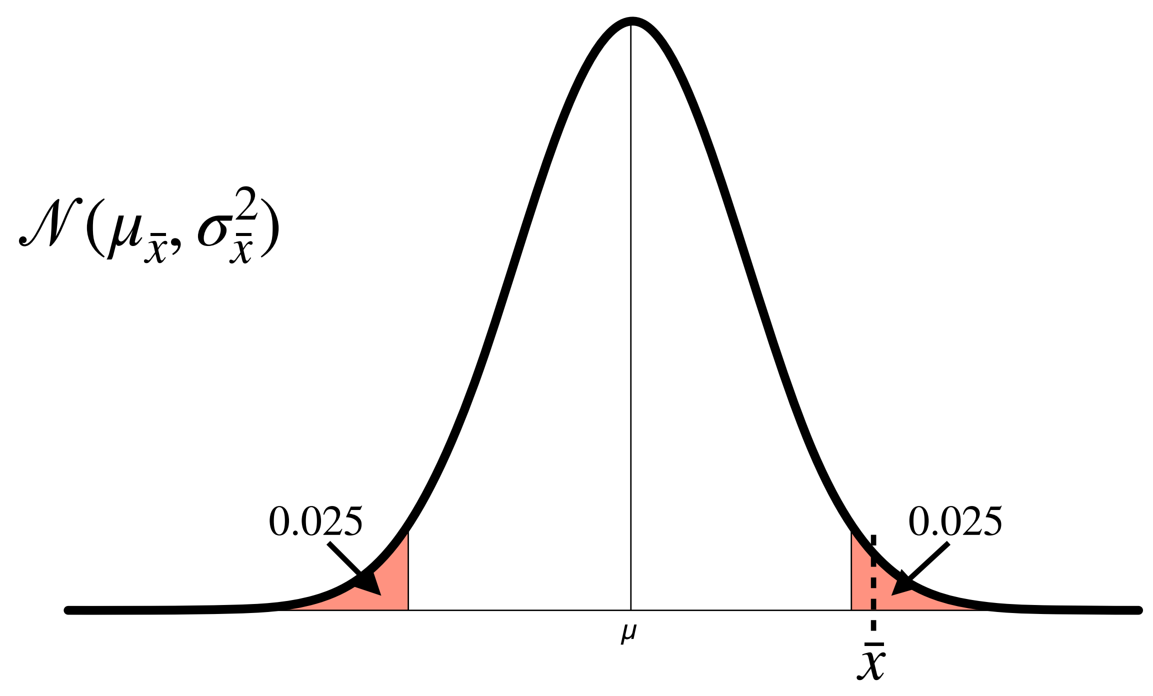
```
import scipy.stats as stats

# compute z = (x − mu) / SE

p = 2 * stats.norm.cdf(−abs(z))
```

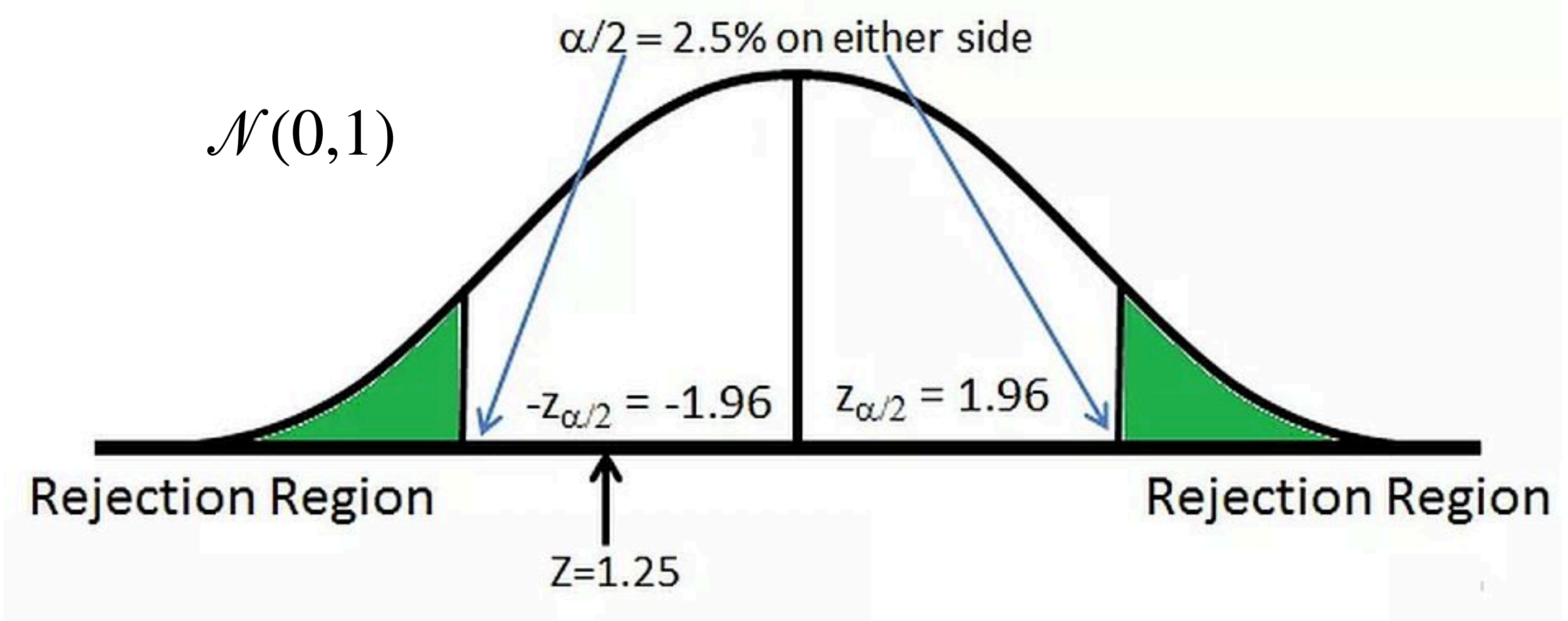
- Why -abs(z)? cdf considers left of the z point, so if z is positive, we want to reference -z

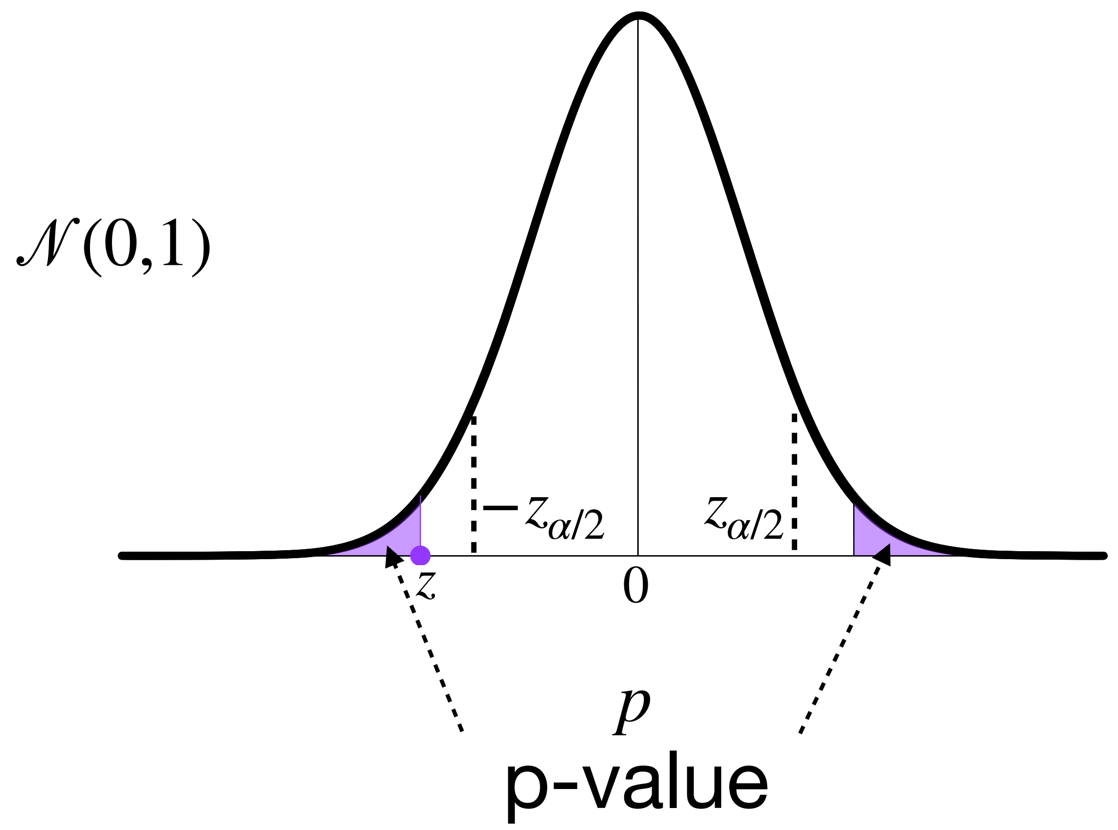
overview of z-test

- Assumptions needed for statistical test
  - Null hypothesis H_0
  - Alternative hypothesis H_1
  - A statistical significance level α
- Equivalent questions (if yes, then reject null hypothesis)
  - Is the sample mean, $\bar{x}$, in tail defined by α of the **sampling distribution** $\approx \mathcal{N}(\mu_{\bar{x}}, \sigma_{\bar{x}}^2)$?
  - Is the z-score, $z = \dfrac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \dfrac{\bar{x} - \mu}{SE}$, in the tail defined by α of a **standard normal** $\mathcal{N}(0,1)$?
  - Is the **p-value**, $p = 2F_{\mathcal{N}(0,1)}(-|z|)$, less than α?

Significance level: $\alpha = 0.05$







back to our original example

- $\mu = 100$, $\sigma = 22$

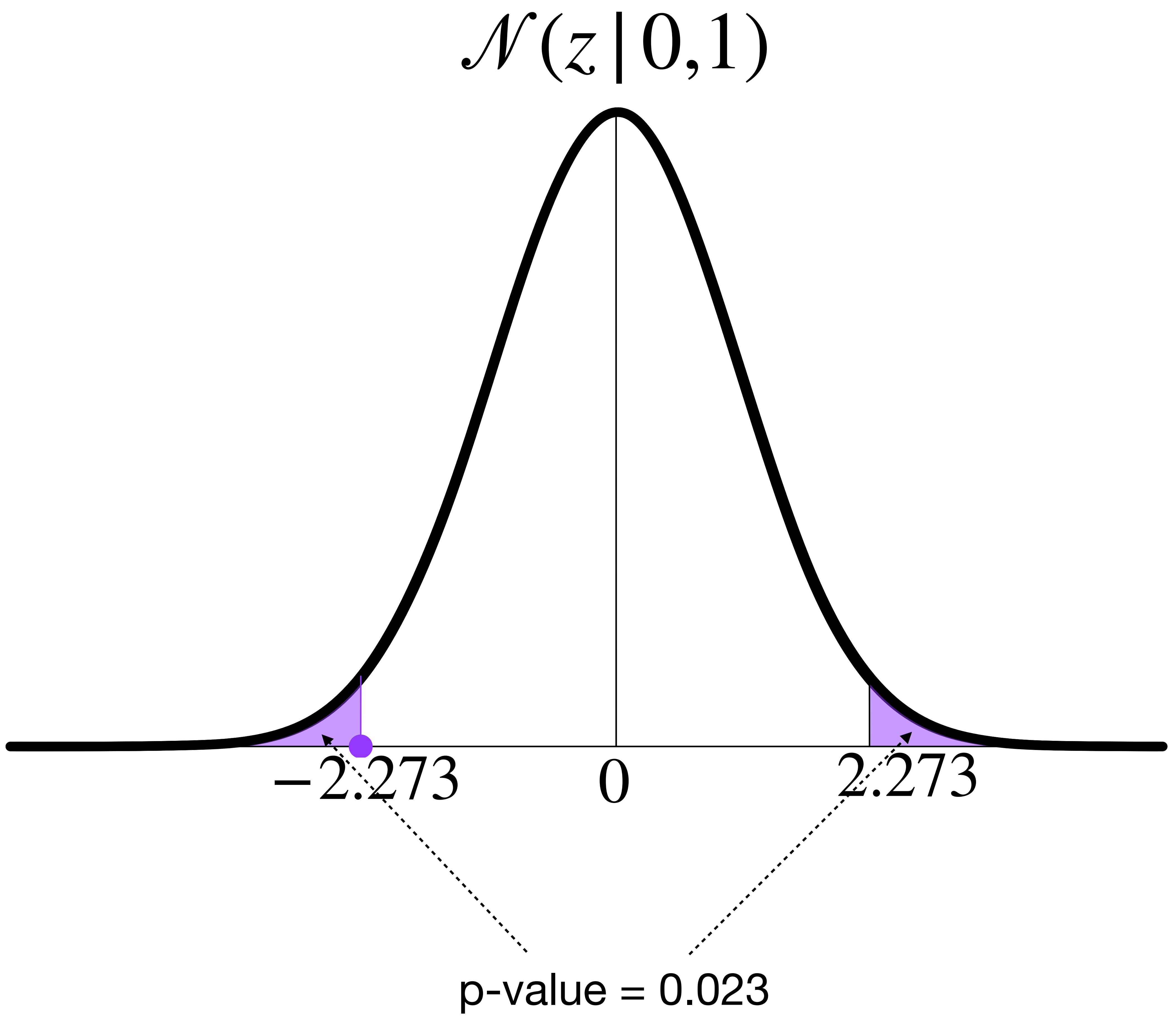
  $\bar{x} = 95$, $n = 100$

- So we calculate:

  $$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{95 - 100}{22/\sqrt{100}} = -2.273$$

  $$p = 2 \cdot F(z \mid 0,1) = 0.023$$

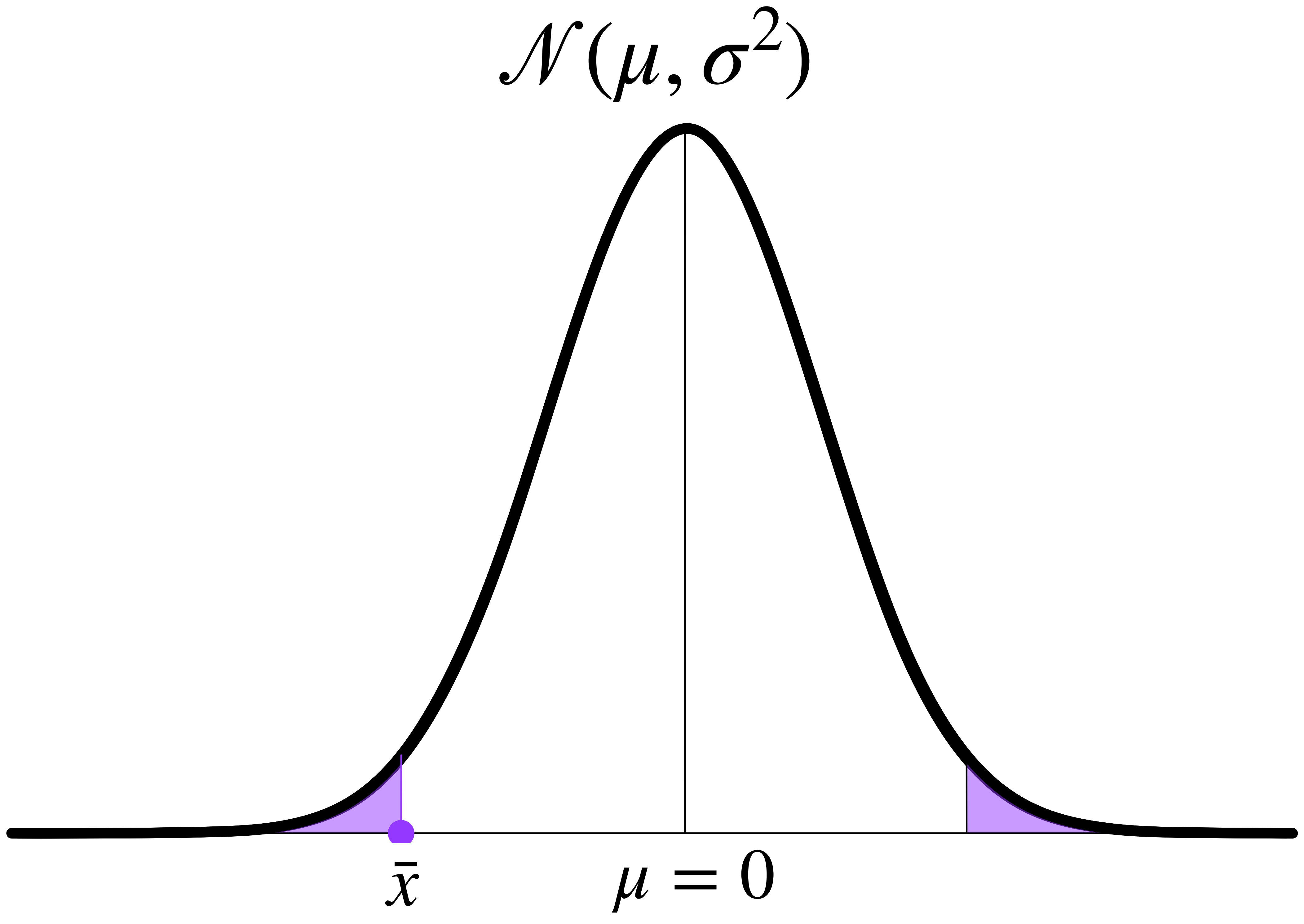
- Conclusion:
  - Significant at $\alpha = 0.1, 0.05$ (reject H_0)
  - Not significant at $\alpha = 0.01$ (cannot reject H_0)



comparing two means

- What if you have *two* populations, and you want to know whether their means are statistically different?
  - Sample 1: Sample size n_0, from pop. mean μ_0, variance σ_0^2
  - Sample 2: Sample size n_1, from pop. mean μ_1, variance σ_1^2
- 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**
- Under null hypothesis, sampling distribution of *difference between two means* has:

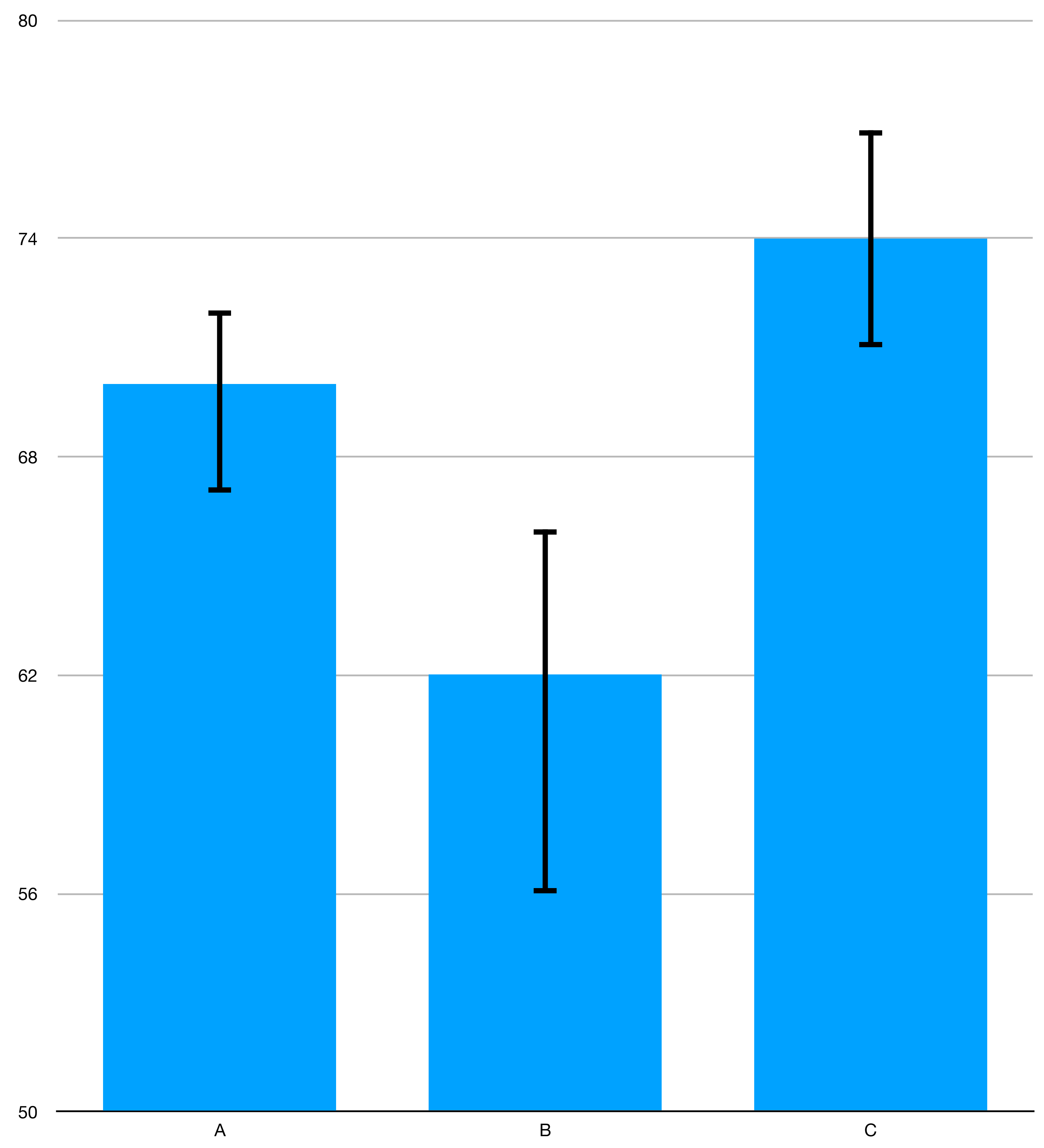
$$\mu = \mu_0 - \mu_1 = 0 \qquad \sigma = \sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}}$$



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

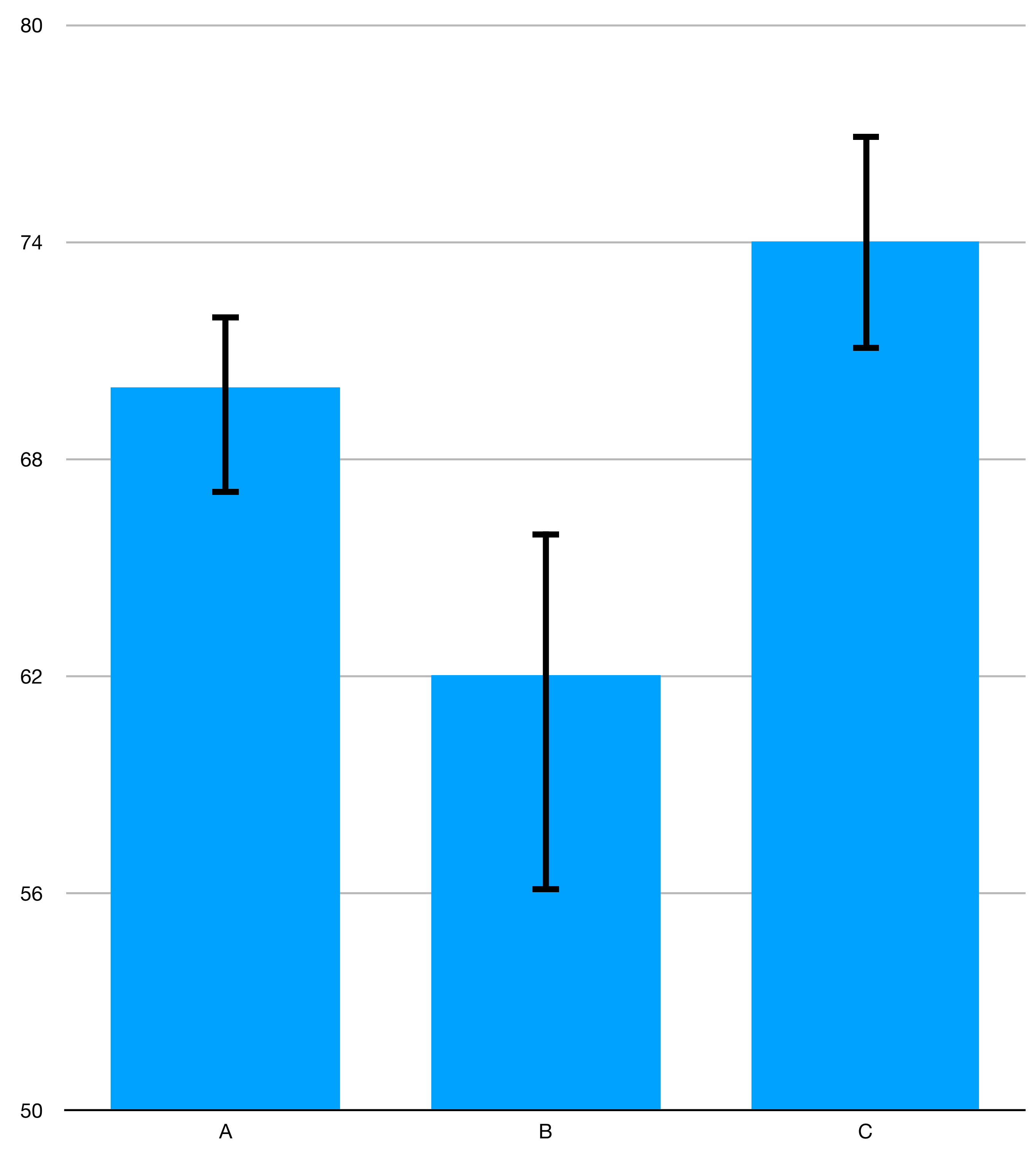
confidence intervals

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



intuition of 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 (95%) confidence interval:
  - *if I were to repeat the experiment a large number of times, 95 percent of confidence intervals would contain the population mean*
  - *before I run the experiment, there is a 95 percent chance that the population mean will fall within the computed confidence interval*
  - *if the population mean is inside the confidence interval, it would not be statistically significant* (informally, you wouldn’t be surprised!)

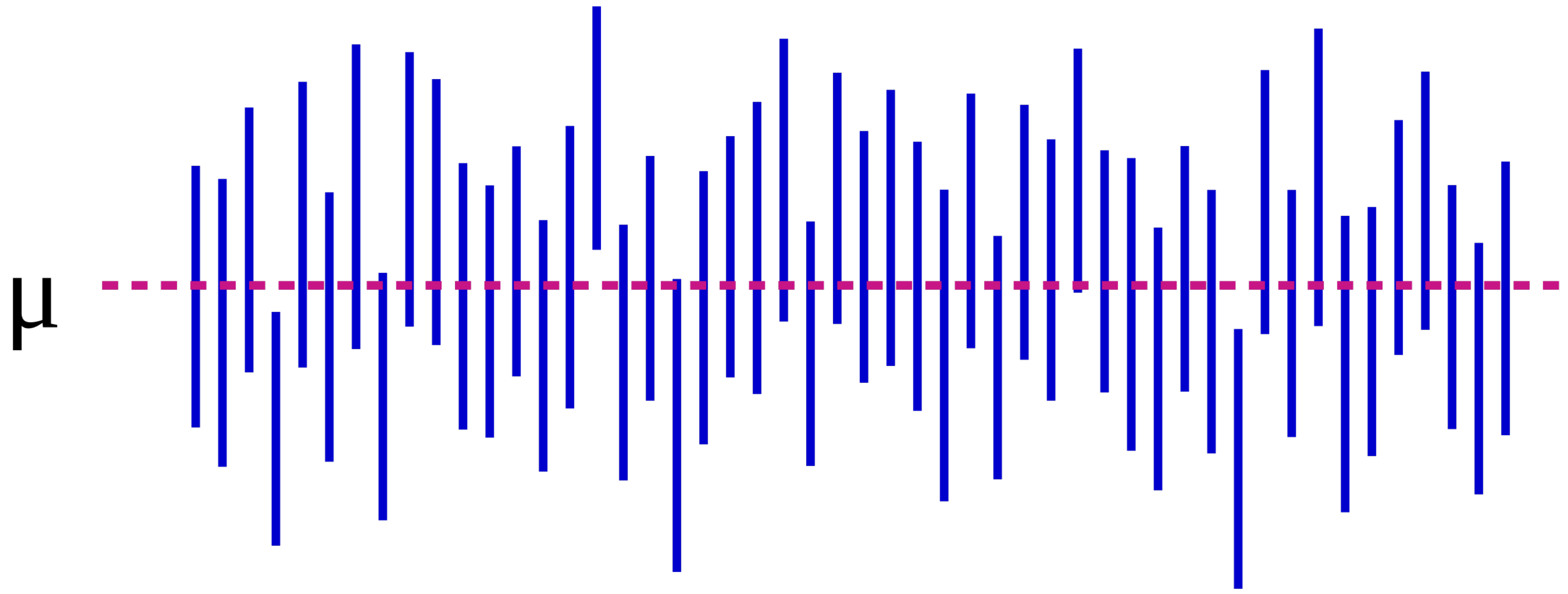


the first interpretation

- *If I were to repeat the experiment a large number of times, 95 percent of confidence intervals would contain the population mean*

- In the diagram below, each vertical bar is one confidence interval calculated for one experiment

    - For a 95% confidence interval, we expect 95% of them will include μ



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

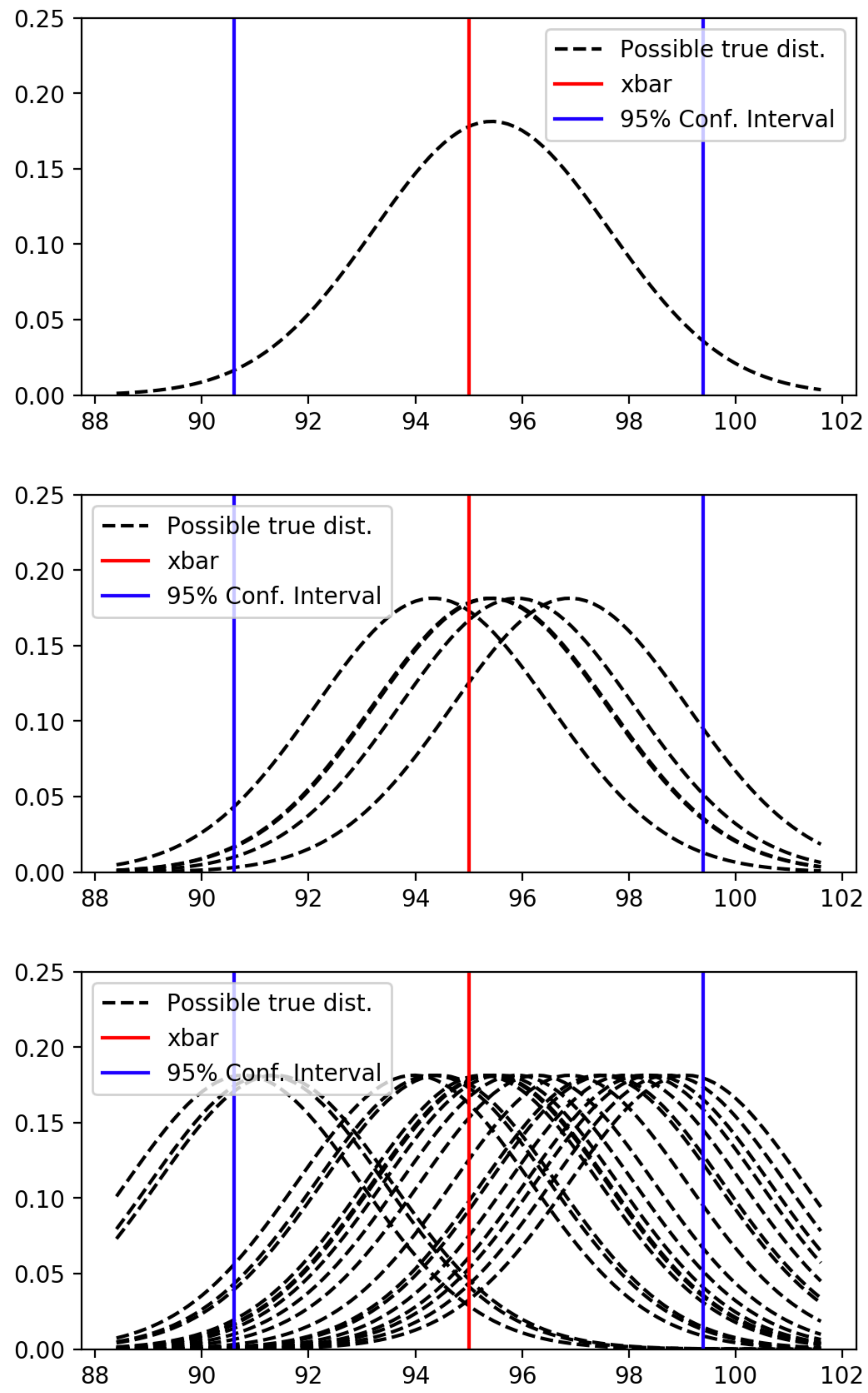
confidence intervals more formally

- 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

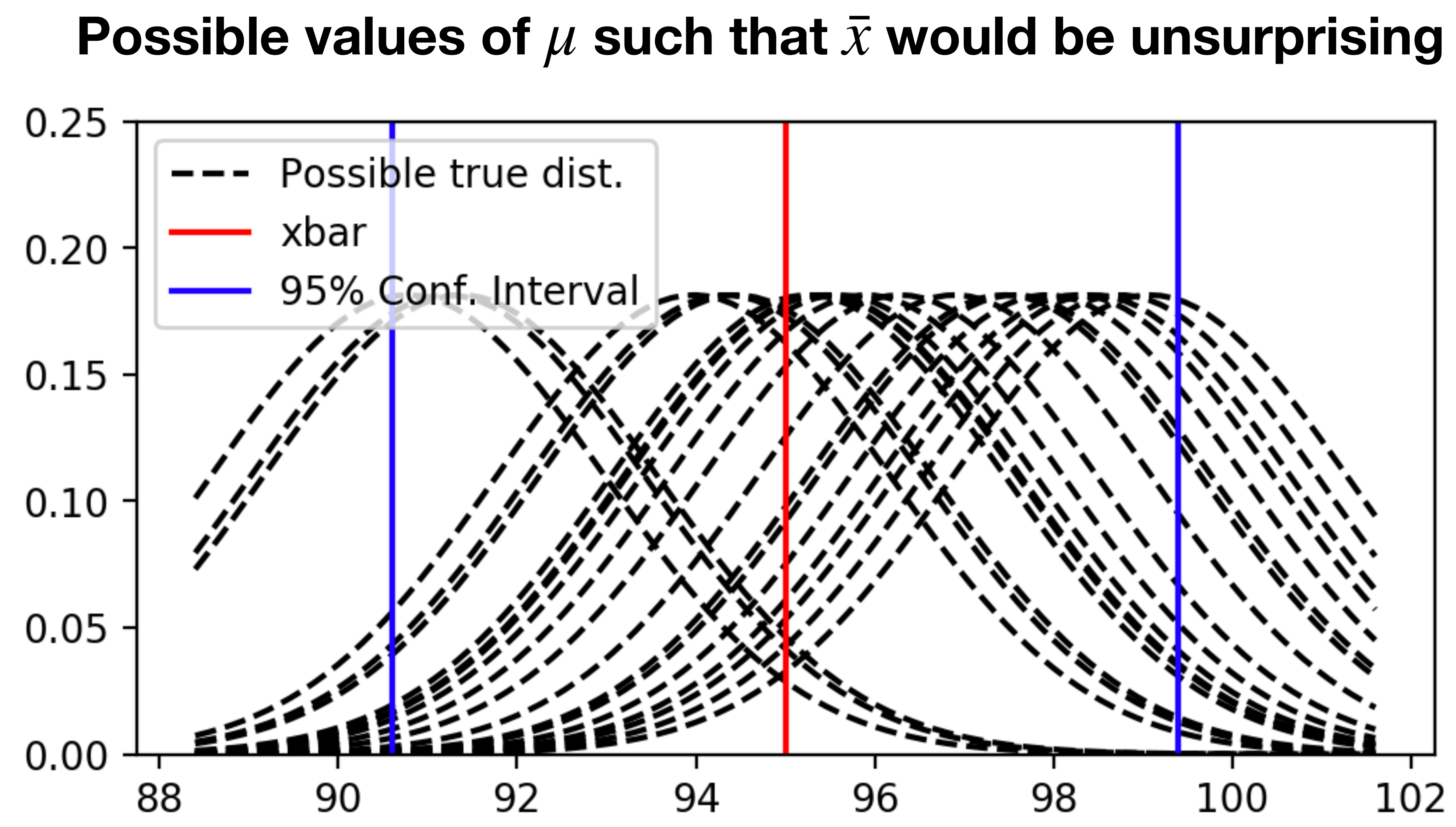
- Conceptually related to z-tests, but the perspective is *reversed*
    - For what sampling distributions (centered at the population mean), would our sample mean NOT be surprising?
    - Note: Our confidence interval is centered around the *sample mean* (instead of the hypothesized population mean)
- Remember definition of z-score:

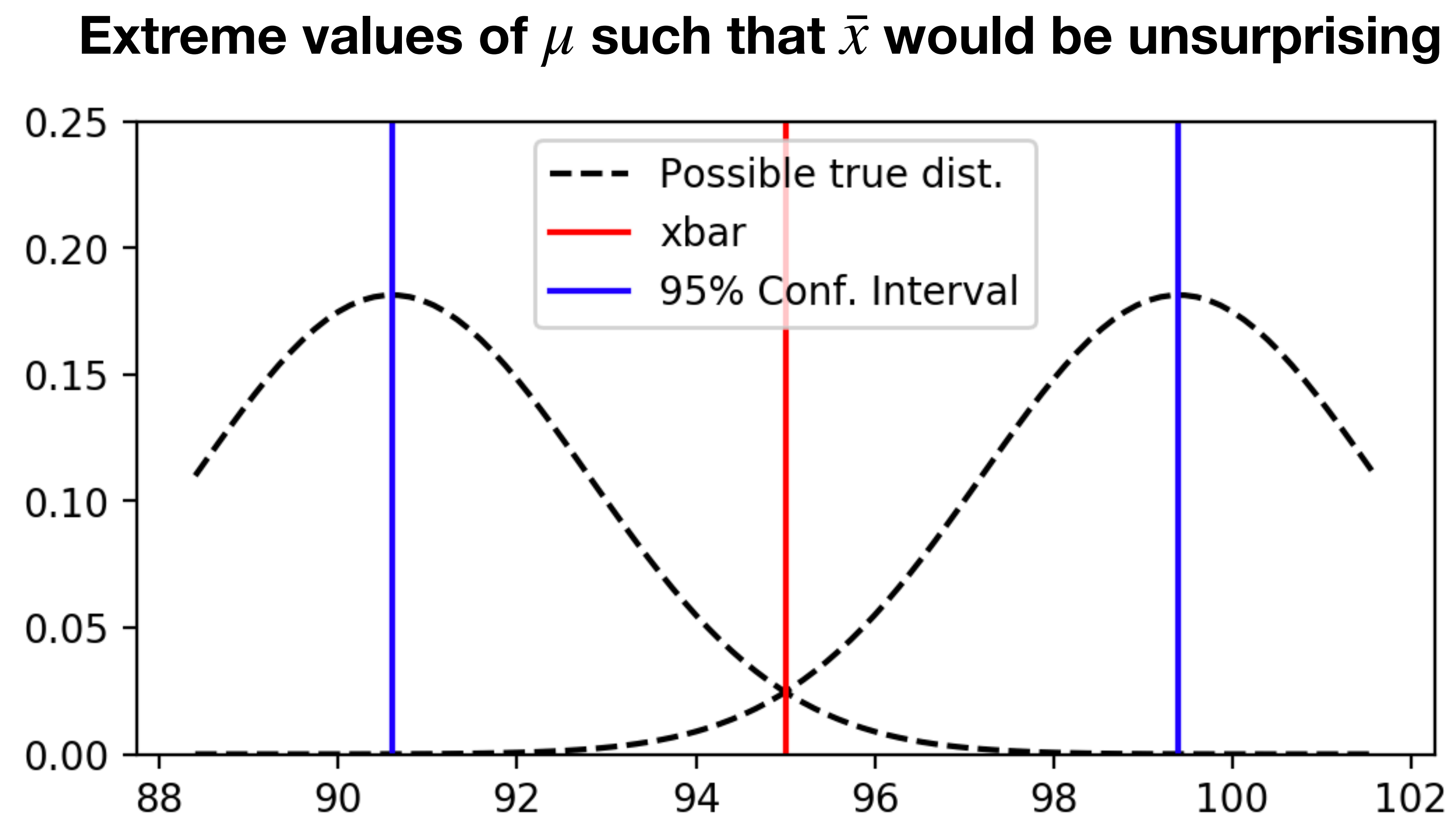
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

- And p-value:

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

- If c is the desired confidence level (here in decimal form), what z do we need such that $p \leq (1 - c)$?





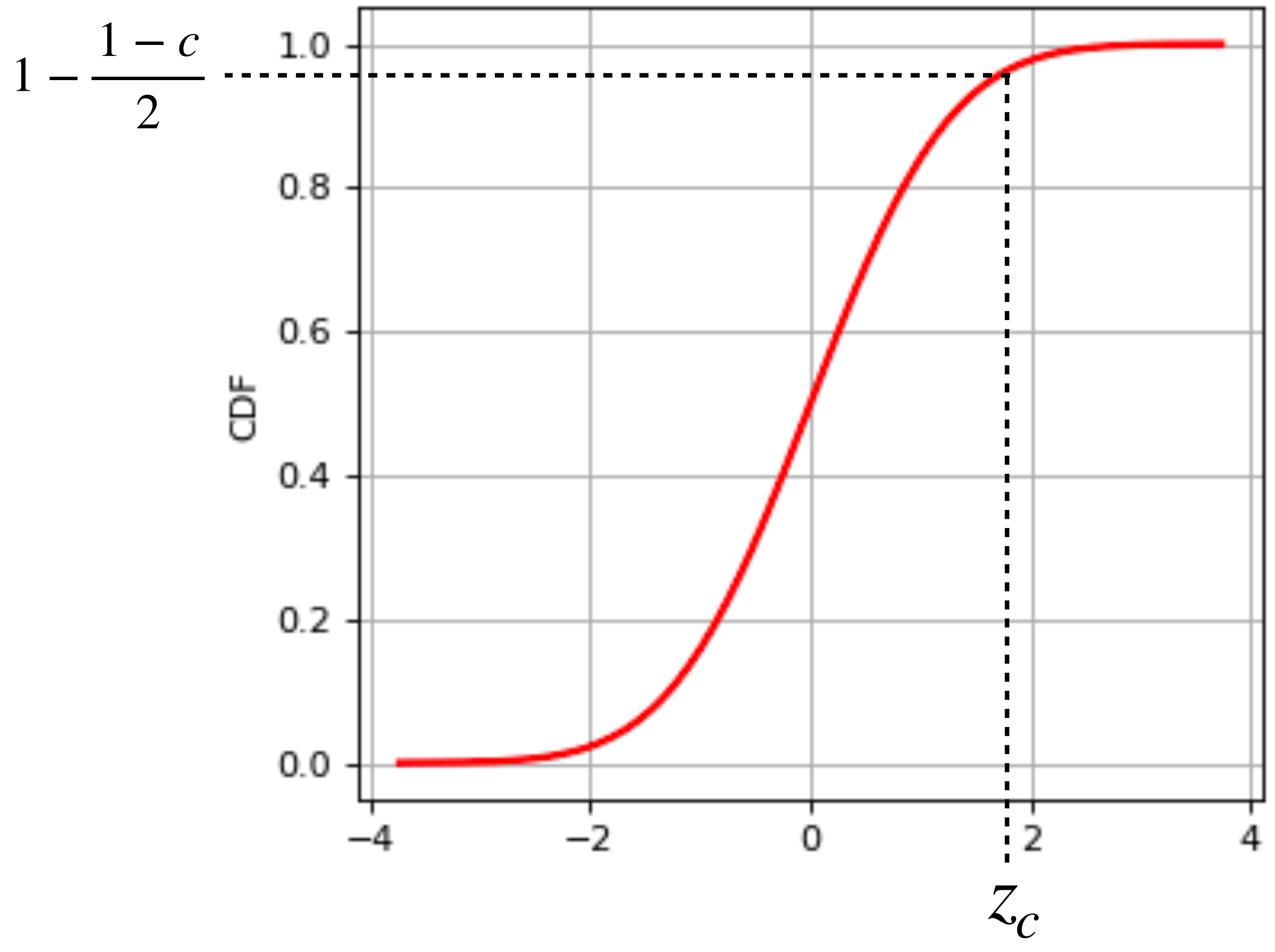
computing confidence intervals

- Call this z_c

- Compute in Python as follows:

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

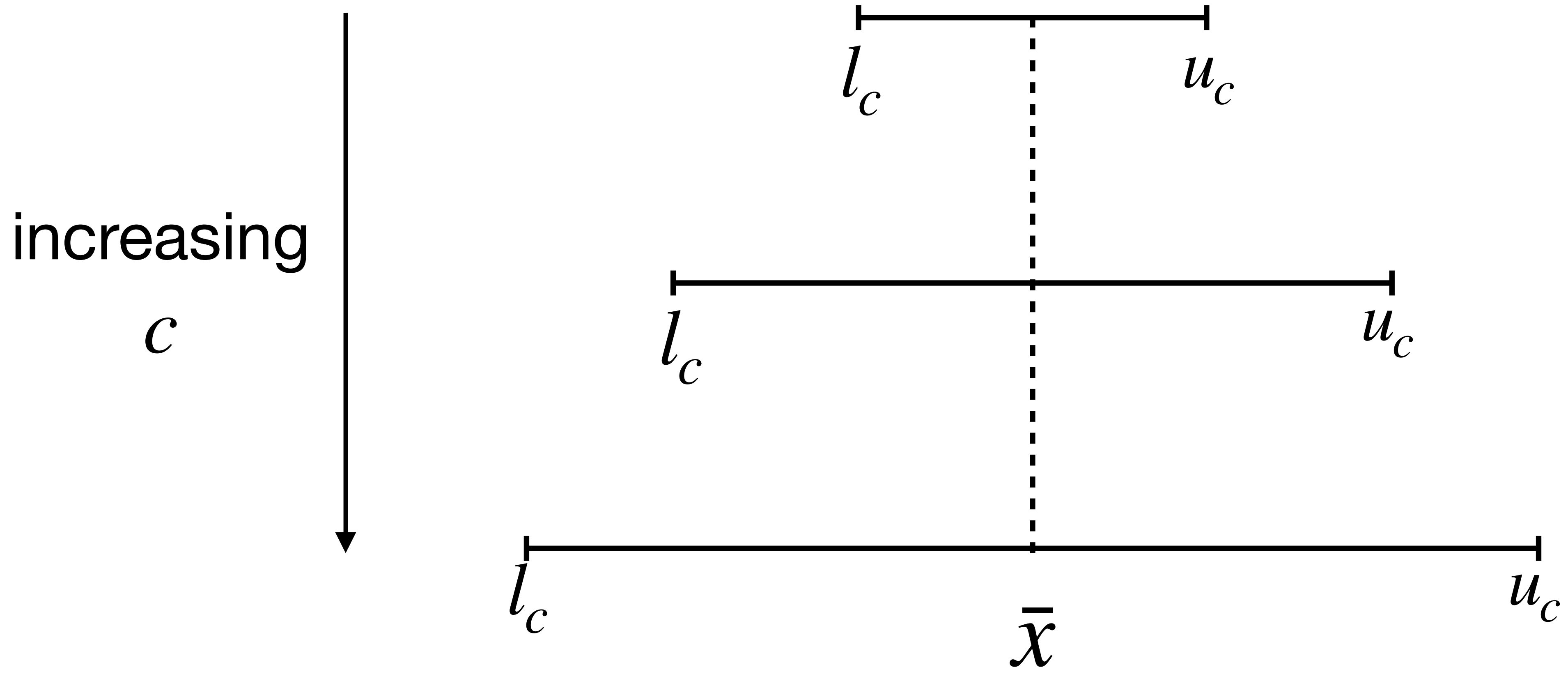
  - While norm.cdf goes from z-score to probability, norm.ppf goes from probability to z-score



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

$$z_c = \left| \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \right| \;\rightarrow\; \mu \in \left(\bar{x} - \frac{z_c \cdot \sigma}{\sqrt{n}}, \bar{x} + \frac{z_c \cdot \sigma}{\sqrt{n}} \right) = (l_c, u_c)$$

  - This is your c% confidence interval



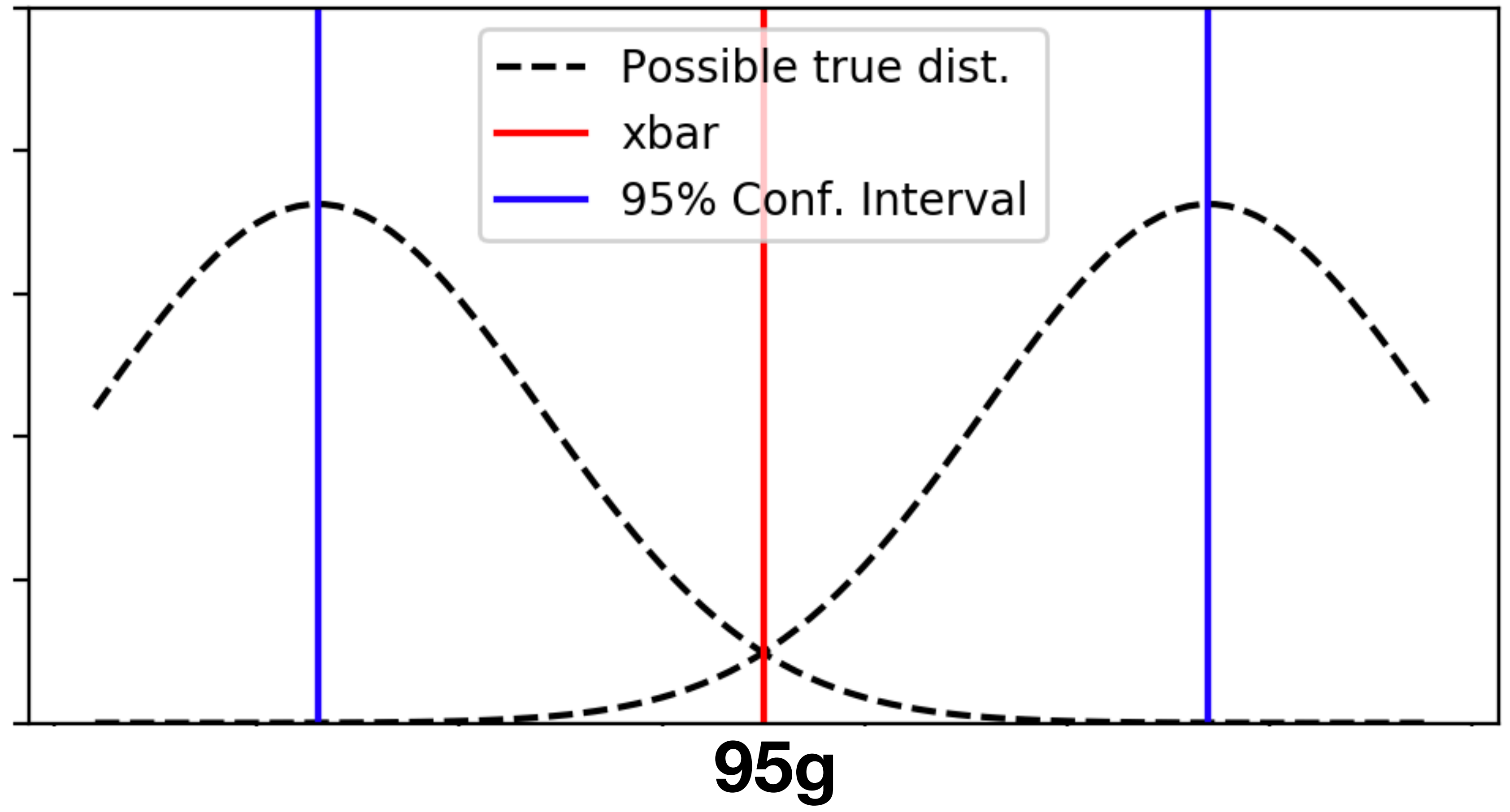
back to our original example

- Let's calculate 90%, 95%, and 99% confidence intervals for μ

- Recall that our sample had

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

- Thus, the confidence intervals are:



$$\mu \in \left(95 - \frac{\sigma}{\sqrt{n}} \cdot z_c, 95 + \frac{\sigma}{\sqrt{n}} \cdot z_c\right) = (95 - 2.2 \cdot z_c, 95 + 2.2 \cdot z_c)$$

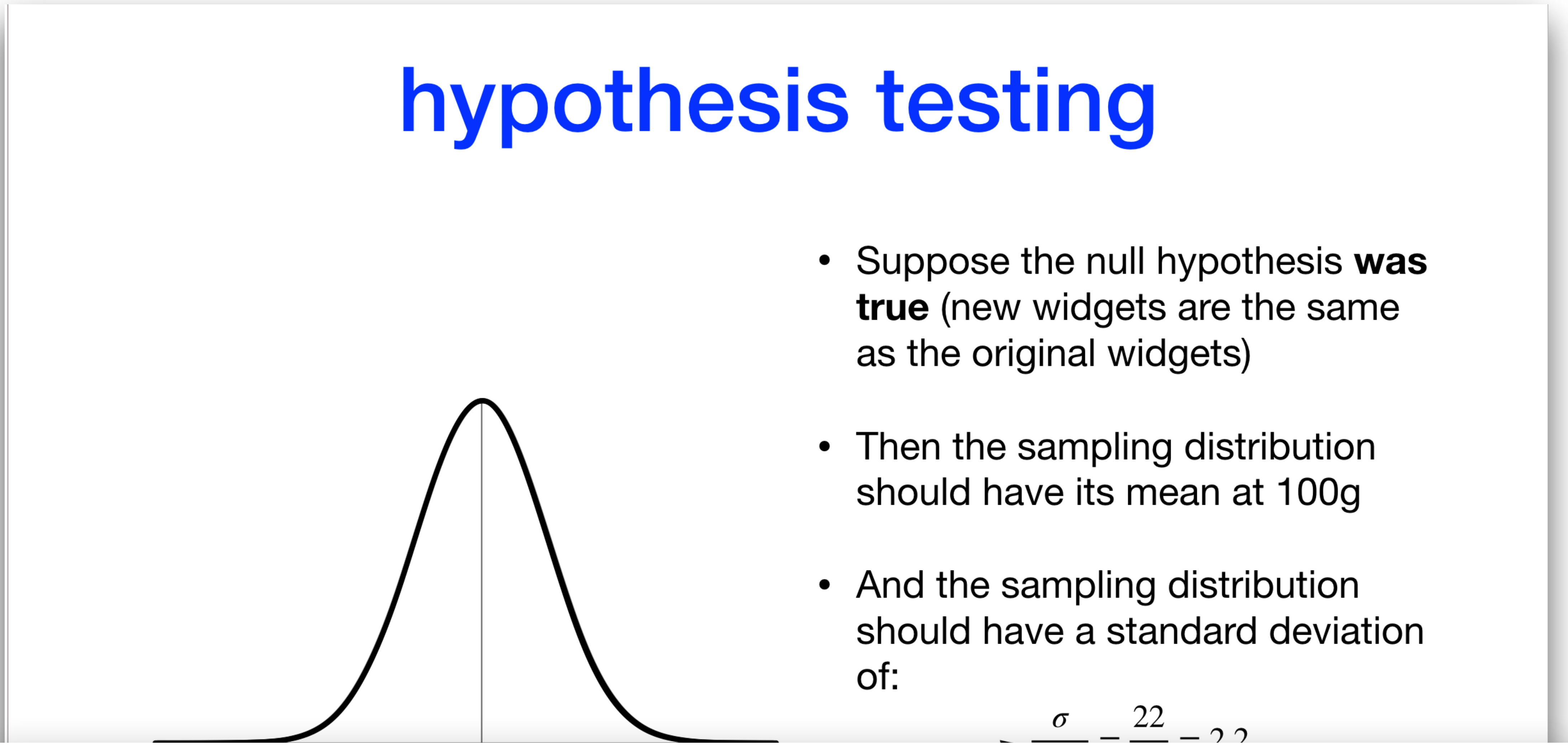
- For 90%, 95%, 99%, $z_c = 1.645,\ 1.960,\ 2.576$. Thus,

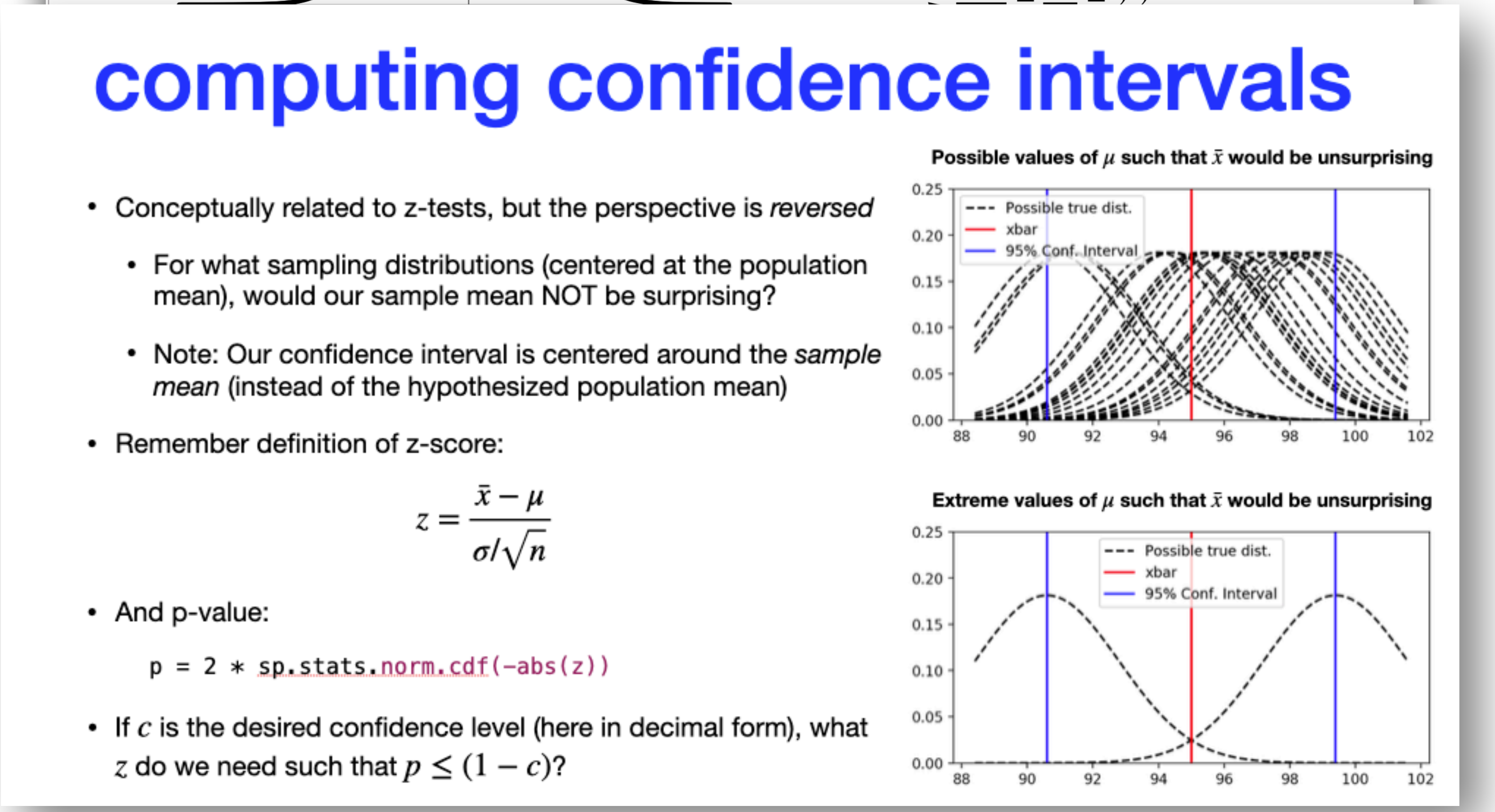
$$90\,\% : (91.38,\ 98.62)$$
$$95\,\% : (90.69,\ 99.31)$$
$$99\,\% : (89.33,\ 100.67)$$

How would we make the intervals narrower for the same levels of confidence?

we’ve been fudging

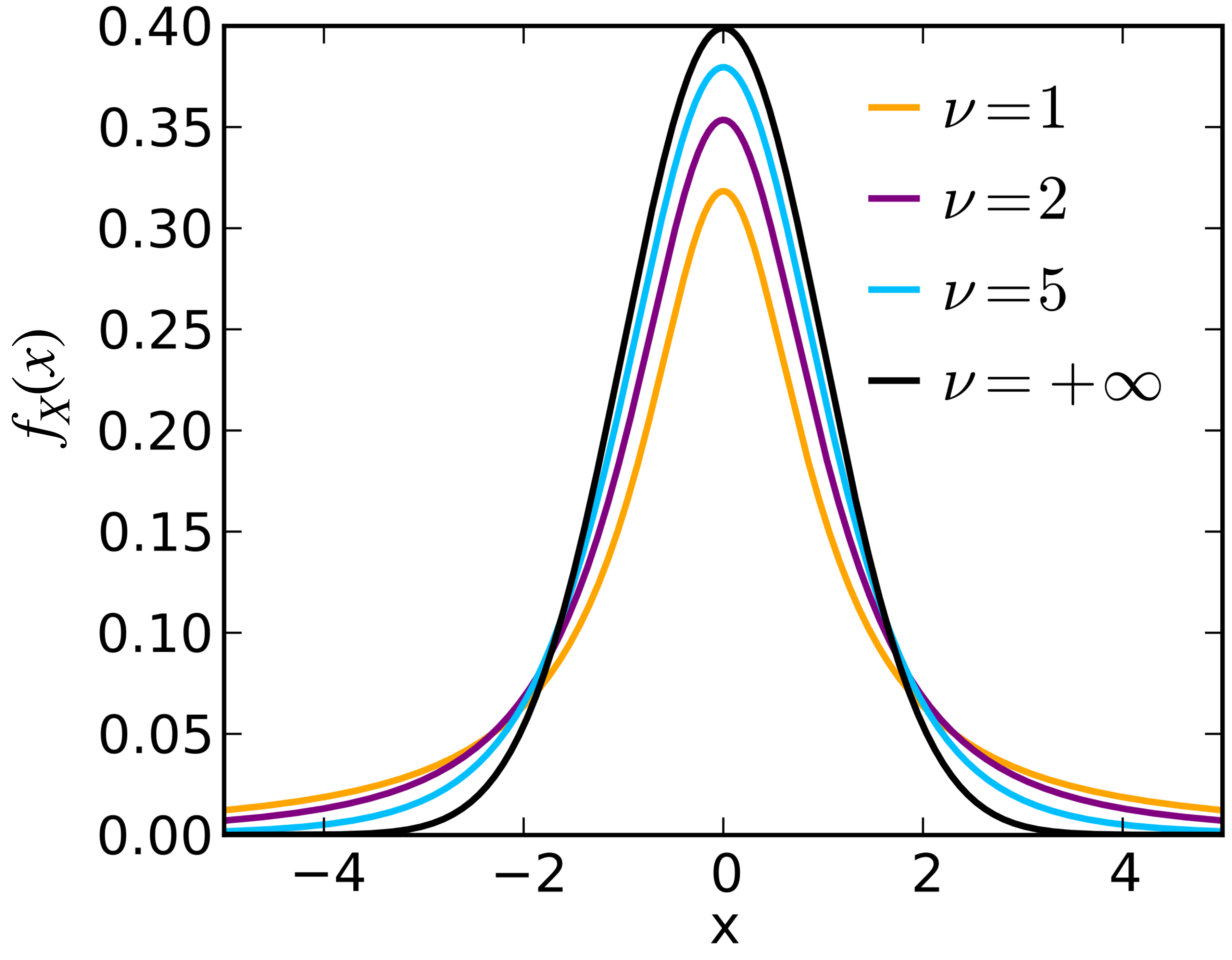
- Recall that to use the z-distribution, we must either know σ or have large enough n
- The **student’s t-distribution** and **t-test** is used when the normal approximation does not hold:
  - i.e., when we don’t know σ (which we usually do not) **<u>and</u>** when $n < 30$
  - Can use this to reason about μ, including building confidence intervals and conducting hypothesis tests





student's t-distribution

- Similar to the standard $\mathcal{N}(0,1)$ normal distribution (density shown to the right)
  - 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 of 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 and confidence intervals

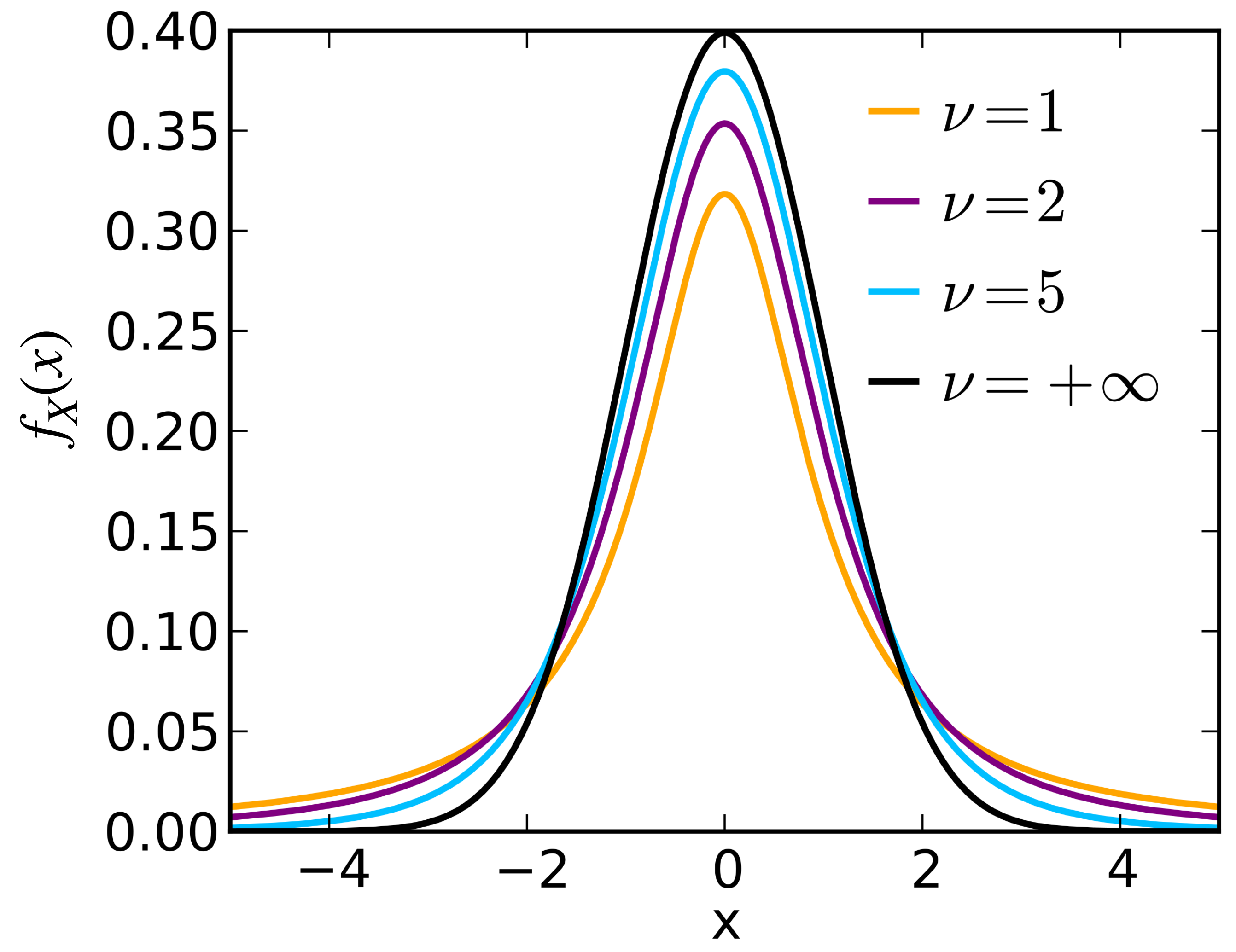
- Works the same as the z-test, except
  - Use s instead of σ
  - Compare to the t-distribution
- Computing the test statistic:
  - First get the standard deviation of the sample:

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

  - Then we get the “t-score”:

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

Compare to the formula for z



- Then we get the p-value:

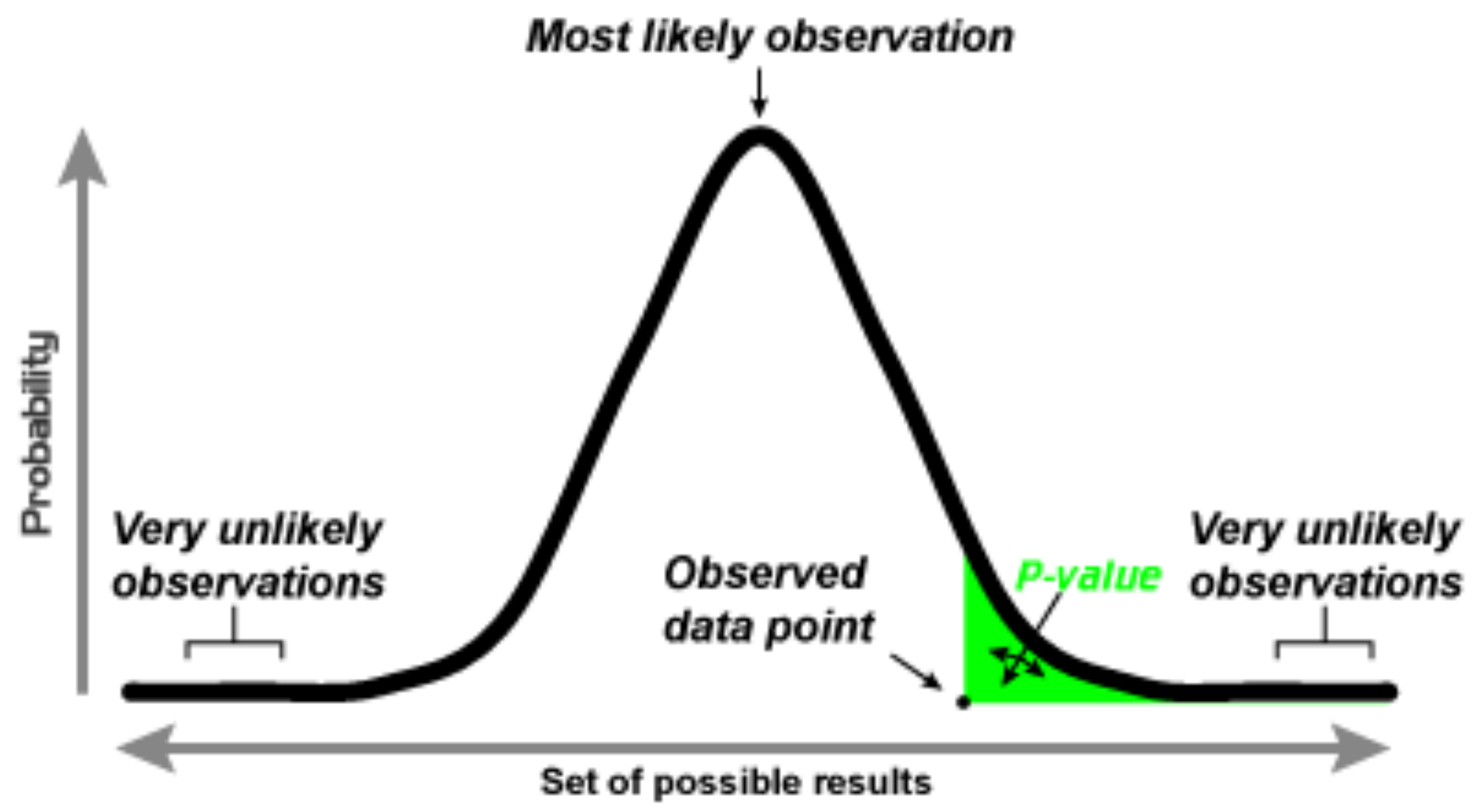
```
p = 2 * stats.t.cdf(-abs(t), df)
```

- And for confidence intervals, we find the t-score corresponding to c:

```
t_c = 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?



- Any given datapoint has *half* the p-value in a one-sided test than it does in a two-sided test
- We also do not divide α by 2 for a one-sided test, because all the area is now in one tail

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, as discussed, 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 $+\infty$
    - When computing z_c or t_c, instead of $1-(1-c)/2$ (where dividing by 2), use $1-(1-c)=c$ since there is only one tail