Lecture 4. Review of Hypothesis Testing, Probability and Inference

Making conclusions about the population from a sample

Nic Rivers

Advanced quantitative research methods, API6319
Fall 2019

Refresher: Mean and standard deviation

- We often want to summarize data.
- For continuous data (or dummy variables):

mean a measure of central tendency. In R, `mean(variable)` or `summary(data.frame)`
$\bar{x} = \frac{1}{N} \sum_i x_i$

standard deviation a measure of variability. The "standard" (or typical) amount of deviation of an observation from the mean. In R, `sd(variable)`.
$sd = \sigma = \sqrt{\frac{1}{N} \sum_i (x_i - \bar{x})^2}$

Our research questions relate to a **population** of interest . . .

- We are typically interested in measuring some phenomenon in a **population**:
  - Univariate phenomenon:
    - What is the unemployment rate of the adult population in Canada?
    - What is the proportion of households that are malnourished in Sub-Saharan Africa?
  - Relationship between two or more variables (multivariate):
    - Does more education lead to higher income (how much more income does an additional year of education cause)?
    - Does higher campaign spending lead to a higher probability of election (how much)?

. . . but we typically do not have information about an entire population.

- One way to answer our research question would be to conduct a **census**.[1]
- For example, to find out about the unemployment rate, we could ask everyone in Canada about employment and job search. We would then have an unequivocal answer to our research question about the unemployment rate.
- However, conducting a census is time-consuming, intrusive, and expensive (e.g., Canadian census costs almost $1 billion to implement.).
- As a result, we don't normally observe the entire **population**, but instead observe only a smaller **sample** of the total population.
  - The Labour Force Survey – which is used to estimate the unemployment rate – is a monthly sample of about 50,000 Canadian households (about 0.2% of the population).
  - Election polls – which are used to predict voting behaviour – typically include samples of 500 – 2,000 people.

[1] A census is a complete enumeration of the population.

We use information from a **sample** to draw conclusions about the **population** of interest

- One of the main goals of statistical analysis is to draw conclusions about a **population** based on measurement from a **sample** of that population.
- This is called **statistical inference**.
- Because we only observe a **sample** of the total **population**, there is **uncertainty** associated with any conclusions that we draw about the population.
- When we make inferences about the population based on a sample, we can't be completely confident in our conclusions.
- When we conduct **statistical inference**, we will try to figure out just how confident we can be in making conclusions about the population.

Random sampling allows us to draw conclusions about the population (mean) . . .

- Suppose the true unemployment rate in the population is 10% (this is normally unobserved; it is an assumption here)
- We will conduct a survey based on a random sample of the population to estimate the unemployment rate
- Imagine **randomly sampling** 100 people from the population and asking them about their employment and whether they are actively looking for a job.
- We can use the information from this random sample to estimate the unemployment rate in the population. For example, if 10% of people in our sample are unemployed, our best guess is that 10% of the population is unemployed.

. . . but there is error associated with a random sample.

- However, we might not be surprised if in our sample, there were 9 or 11 people (or even 12 or 13 people), because we know that there is some randomness: we might have (randomly) happened upon a group of people that were more or less likely to be unemployed than the average.
- If, for example, our sample of 100 people contained only 8 people who were unemployed, we would erroneously conclude that the unemployment rate in the population was only 8%.
- Any time we draw a conclusion about the population based on a sample, there is a chance we will make an error.

Standard Error

- The standard error is an estimate of how much error is typically associated with estimating some characteristic of the population (typically the mean) from a sample of the population. It is the "standard" amount of error we expect in extrapolating from a sample to a population.
- Formally, the standard error is the standard deviation of the distribution of possible sampling outcomes.
- Note the difference between the standard error and standard deviation:

  Standard deviation Captures the typical deviation between an observation and the sample mean.

  Standard error (of the mean) Captures the typical deviation between the sample mean and population mean.
- Where does the error come from in a random sample?
  - From the randomness. There is a chance that we will randomly choose a sample that has different characteristics than the population.

The standard error and the standard deviation

- The **standard deviation** (σ) expresses an estimate of the average deviation of an observation from the sample mean.
- The **standard error** (se) expresses the average deviation of the estimate of the sample mean from the population mean.
- How are the two related?
  - The standard error of an estimate of the mean is:

$$se(\bar{x}) = \frac{\sigma}{\sqrt{N}}$$

  - The standard error of the mean will be smaller if the standard deviation is smaller
  - The standard error of the mean will be smaller if the number of observations in the sample is larger
  - Note that the standard error gets smaller (our estimate of the mean gets more precise) as the size of the sample gets bigger; the standard deviation does not.

Unemployment example

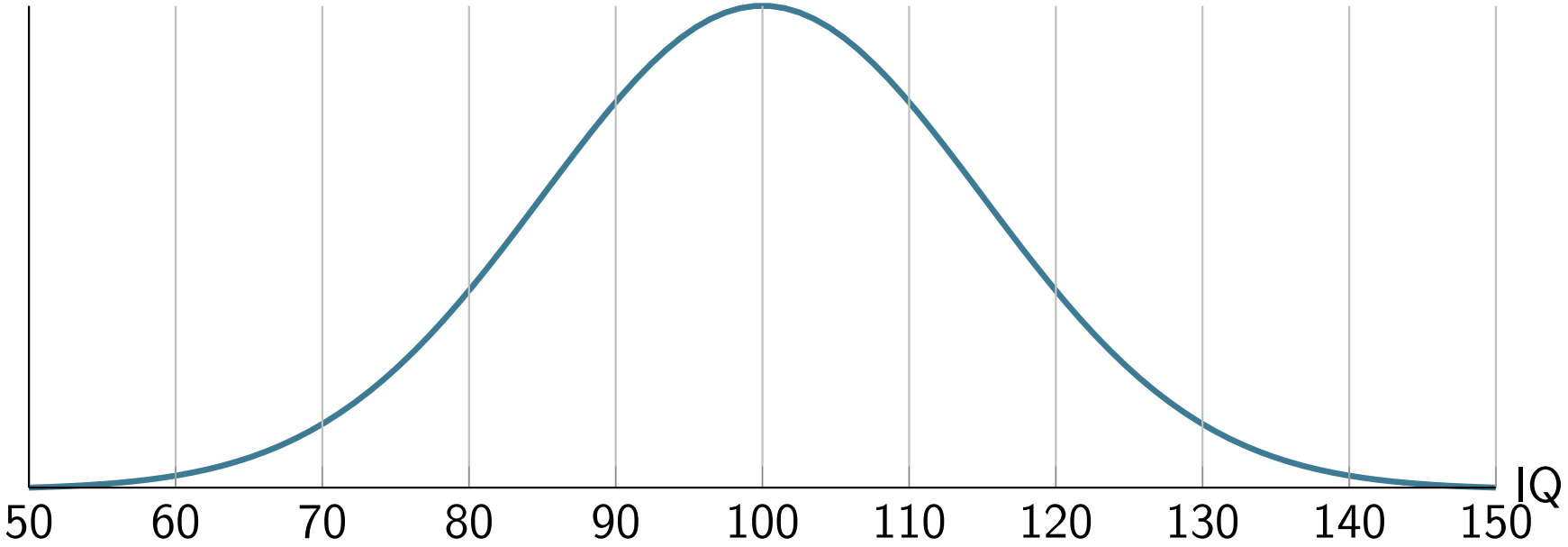
- Imagine we randomly sample 100 people, and find that 11 of them are unemployed. What can we say about the unemployment rate in the population?
  - Think of a dummy variable `unemp` that takes on a value of 1 if the person is unemployed and 0 if the person is employed
- We can calculate:
  - Mean unemployment in the sample: $\bar{x} = \frac{1}{N}\sum_i x_i = 0.11 (= 11\%)$
  - Standard deviation: $SD = \sqrt{\frac{1}{N}\sum_i (x_i - \bar{x})^2} = 0.315 (= 31.5\%)$
- We know that there is uncertainty involved in making conclusions about the population based on observing just a sample of the population.
- Because of random variability, we define the **standard error** as the chance error in the estimate of the mean
- Standard error is $SE = \frac{sd}{\sqrt{N}} = \frac{0.315}{\sqrt{100}} = 0.0315 (= 3.15\%)$
- The *standard* amount of error in our estimate of the mean is 0.0315 (or 3.15%). (Our best guess is that mean of unemp is 0.11. But we expect to be wrong by about 0.032 in our conclusion.)

Sources of potential errors

- Two sources of error:
  1. Sampling error from random sample: Even in a random sample, there will be errors in conclusions because we may inadvertently (randomly) include more of some types of individuals their share in the population.
  2. Bias from non-random sample: If we do not choose a random/representative sample, we will make errors extrapolating the results to the population (e.g., landline polls for elections)
- In this class, we focus on sampling error, and assume that we have a representative sample. You need to think carefully about what your sample represents in your own work.

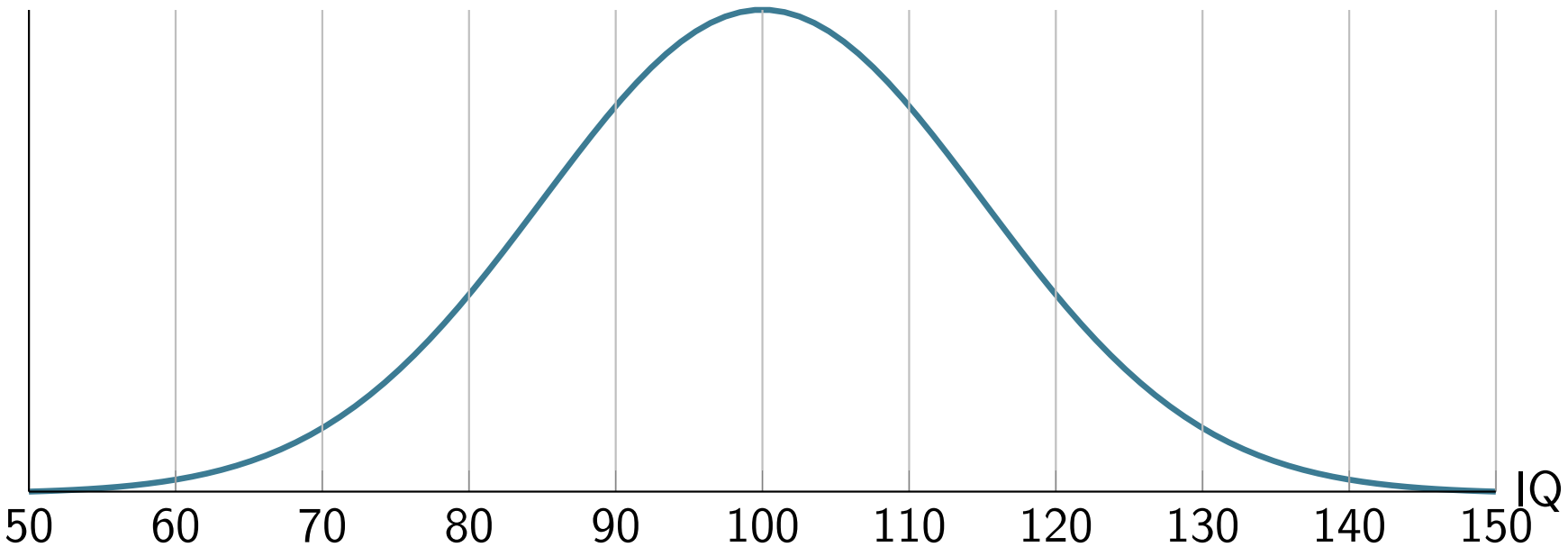
The normal probability density function

- The **normal distribution** is the most widely used and important probability density function in statistics.
- It is sometimes called the "bell curve" or "Gaussian curve"
- Many variables are (approximately) normally distributed (height, weight, IQ)



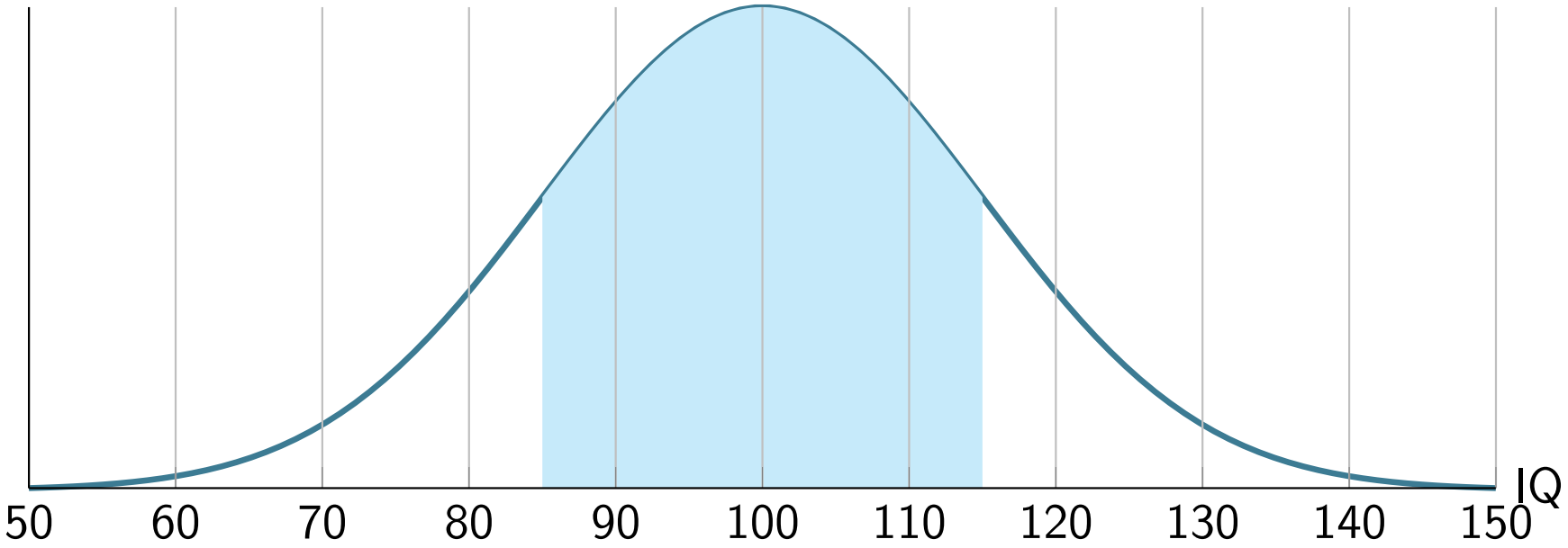
The normal probability density function

- The normal distribution is *symmetric* with mean μ and standard deviation σ.
- We write this in shorthand as $N(\mu, \sigma)$.
- For example, IQ is normally distributed as $N(\mu = 100, \sigma = 15)$.



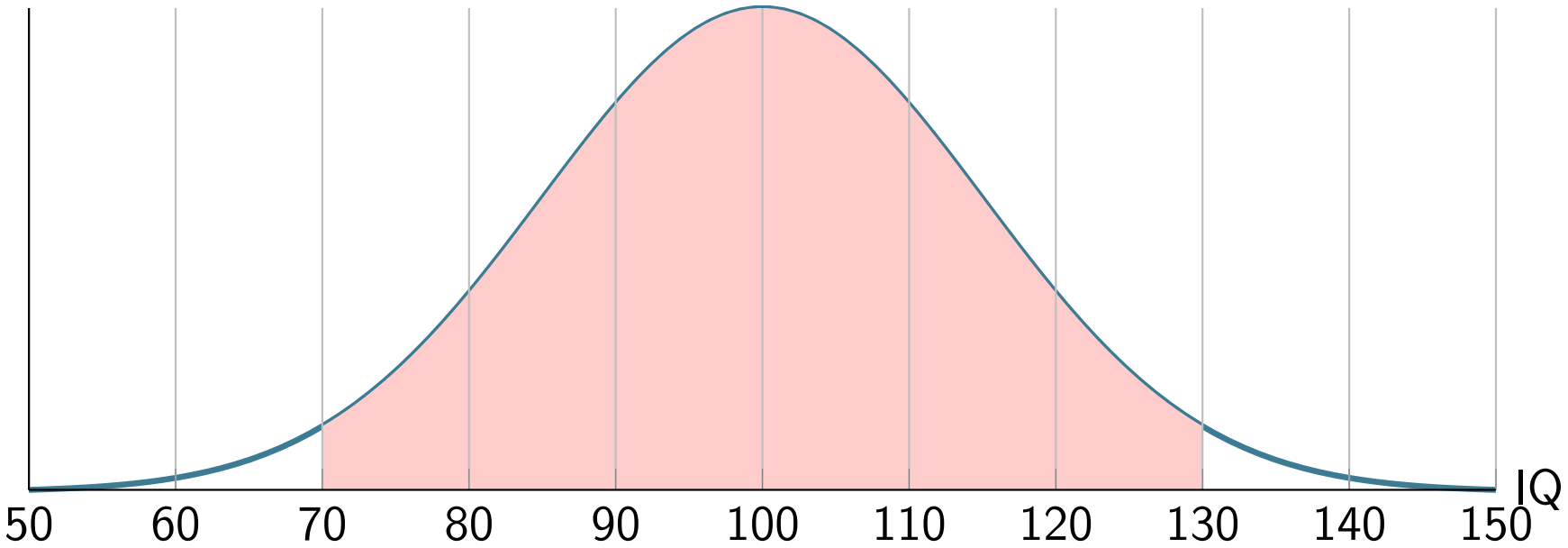
The 68-95-99.7 rule for normal distributions

- In a normal distribution, approximately 68% of the density (i.e., 68% of observations) are within 1 standard deviation from the mean



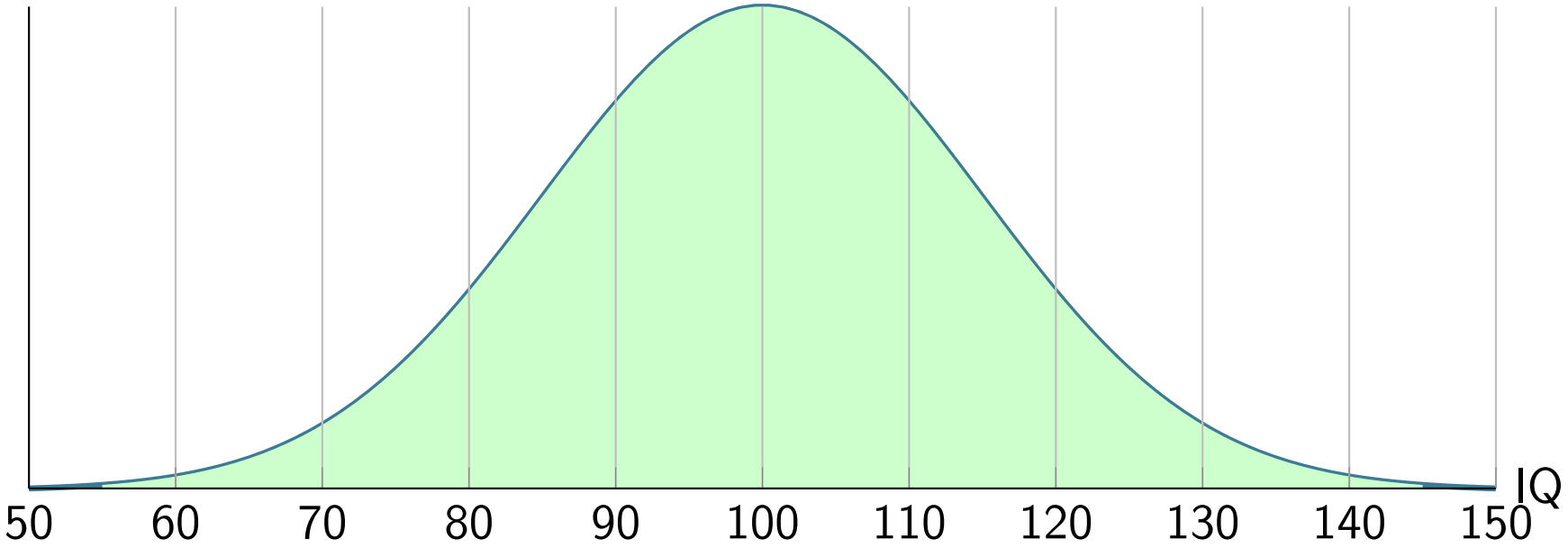
The 68-95-99.7 rule for normal distributions

- In a normal distribution, approximately 95% of the density (i.e., 95% of observations) are within 2 standard deviations from the mean



The 68-95-99.7 rule for normal distributions

- In a normal distribution, approximately 99.7% of the density (i.e., 99.7% of observations) are within 3 standard deviations from the mean



Central limit theorem (1)

- Problem: We know about the normal distribution, but many variables are not normally distributed.
- Example: unemployment is a discrete variable (yes/no), so is clearly not normally distributed.
- Example: per capita income is also not distributed normally.
- How can we use the information we have about the normal distribution to help us in analyzing real variables?

The central limit theorem (2)

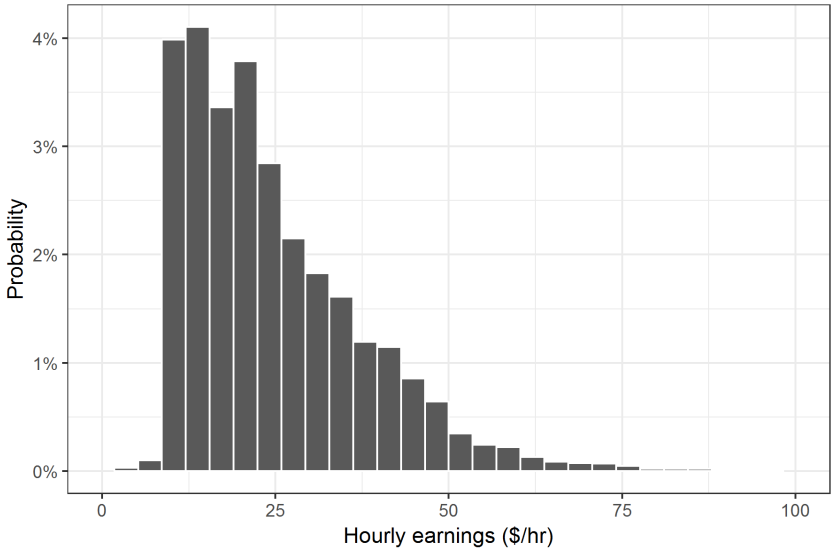
Theorem (Central limit theorem)

For any random variable (with any distribution), if we take repeated samples, and find the average of each sample, these averages will be normally distributed

- For example - think of a coin toss: only takes on values of 0 (heads) or 1 (tails) - clearly not normally distributed.
- BUT, take samples of 100 coin tosses, and count the number of heads. When number of samples gets large, this will be normally distributed.
- So, even where a variable is not normally distributed, an estimate of the mean is generally normally distributed.

Central limit theorem example

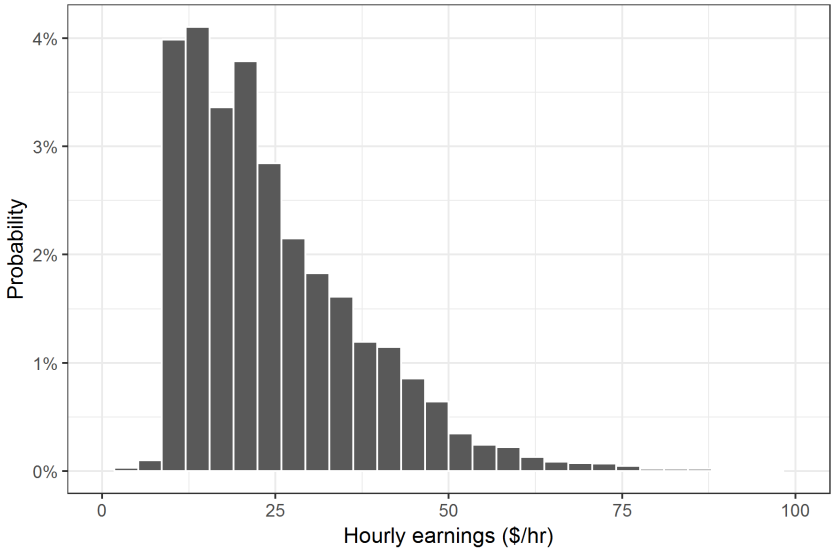
Hourly wages



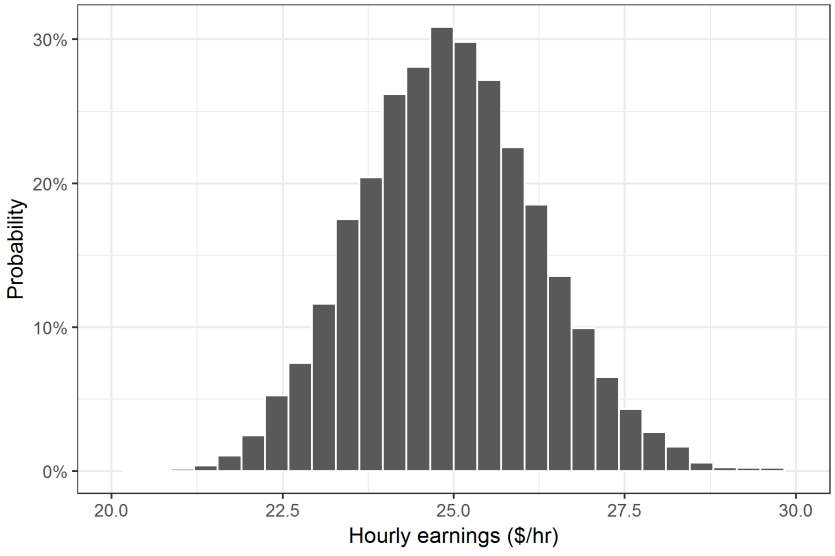
Hourly wages are not distributed normally.

Central limit theorem example

Hourly wages



Hourly wages are not distributed normally.



If we take a sample of people, and calculate the average hourly wage in the sample, then repeat, the **means** from these samples are normally distributed.

The central limit theorem (3)

Using the normal distribution for hypothesis testing

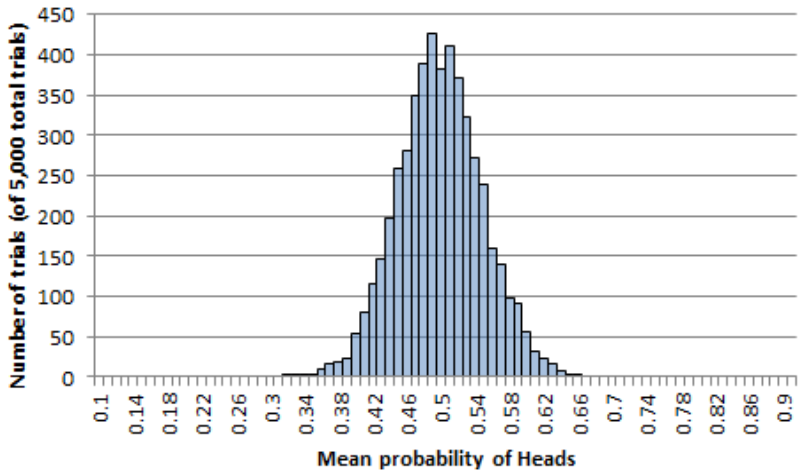
- Imagine your friend has a coin that you suspect is "weighted" (i.e., heads comes up less than 50% of the time).
  - Hypothesis: the coin is weighted (different than a fair coin).
  - Null hypothesis:

The central limit theorem (3)

Using the normal distribution for hypothesis testing

- Imagine your friend has a coin that you suspect is "weighted" (i.e., heads comes up less than 50% of the time).
  - Hypothesis: the coin is weighted (different than a fair coin).
  - Null hypothesis:the coin is not different than a fair coin.
- The data: your friend flips the coin 100 times and tails comes up 30 times.
- Do we have enough evidence to conclude that the coin is not 'fair'?
  - Should we "reject" the null hypothesis?

Figure: Outcomes from sets of 100 flips of a fair coin



The central limit theorem

- What about if heads comes up 40 times?
- What should be our rule for determining if the coin is fair? Note that for any rule, there is a possibility that we will erroneously conclude that the coin is unfair, even if it was actually fair. (it it *possible* — but unlikely — that a perfectly fair coin could produce 100 heads in a row)
- The convention in statistics for hypothesis testing is to use a 5% rule. In other words, if a result falls outside of 95% probability interval, we assume that it is inconsistent with our hypothesis.
- Recall that for a normal distribution, 95 percent of the probability is within 2 standard deviations from the mean.

The central limit theorem (5)

- If we accept a 5% chance of concluding that the coin is unfair when it is actually fair, our threshold should be:

$$
\begin{aligned}
&= mean \pm 2 \times se \\
&= mean \pm 2 \times \frac{\sigma}{\sqrt{N}} \\
&= 0.5 \pm 2 \times \frac{0.5}{\sqrt{100}} \\
&= 0.5 \pm 0.1 \\
&= [0.4 - 0.6]
\end{aligned}
$$

- We would treat coins with 40-60 heads out of 100 as "fair" and others as "unfair".
  - Reject the null hypothesis if outcome is not in this range.
- Note that in our distribution, 5% of the time, the fair coin comes up with values outside this range.

Inferences about the population from a sample

Admitting the chance of mistakes

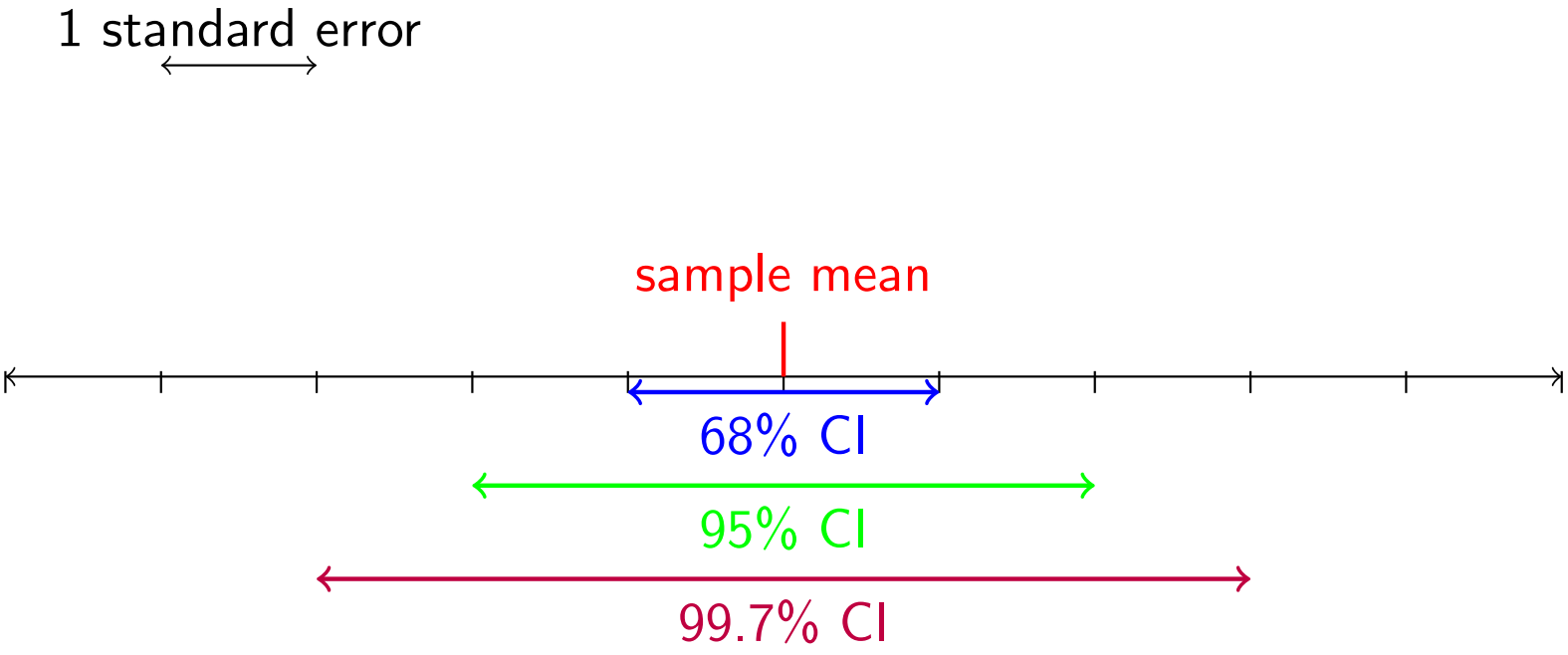
- Any time we make an inference about the population from a sample, we have a chance of making an error
- We are able to quantify how much error we are likely to make
- We can therefore quantify our chance of being wrong when we make an inference about the population
- We call this chance α. α is the probability that we will reject our null hypothesis when it is really true.
- (We declare the coin unfair, even though it's a regular fair coin.)
- There is no absolute rule for what we should choose as α. It might reasonably differ in different circumstances.
- In practice, we often select $\alpha = 0.05$. This means that we admit a 5% chance of rejecting the null hypothesis when it is really true.

Point estimates and standard errors (con't)

- We can use the standard error to generate a confidence interval around our estimate of the mean
- Rule of thumb: a 95% **confidence interval** can be constructed as $(\bar{x} \pm 2 \cdot se)$. (note: this is an approximation; actual values are closer to 1.96).
- A 95% confidence interval is interpreted as follows:

  If we were to collect the same data multiple times, 95% of the time the true population mean would be contained in the constructed confidence intervals.

Confidence interval size



The confidence interval reflects a trade-off between uncertainty and precision. If we want to be sure about not being wrong (e.g., 99.7% CI), we need to admit a wide range of outcomes is possible.

Hypothesis testing

- Normally, our aim in applied statistics is in hypothesis testing
  1. We start with a hypothesis. (The coin is weighted.)
  2. Translate to a null hypothesis. (The coin is no different than a normal coin.)
     - For a null hypothesis relating to the sample mean, this will be in the form of:
     - H_0: the true population mean is μ_0 (The mean rate of heads on the coin is 50%.)
  3. We observe some data from a sample of the population. Based on the sample, we calculate statistics for the sample (such as the mean and standard deviation).
  4. We make a determination of whether the observed data is consistent or inconsistent with the null hypothesis.
     - This involves calculating a **test statistic** for the sample, and using a rule to determine whether it is consistent with the null hypothesis. (30 heads out of 100 is not consistent with a fair coin.)
  5. If the data is inconsistent with the null hypothesis, we reject the null hypothesis.

Statistical test about the sample mean: The **t-test**

- We will use a formal test to determine whether we should reject hypothesis about the sample mean.
- The test involves comparing the null hypothesis (μ_0) to the sample mean ($\bar{x}$).
- Based on the *central limit theorem*, we know that an estimate of the mean is normally distributed
- We can therefore use a normal distribution to test a hypothesis relating to the sample mean
- This statistical test is called the **t-test** and is the most important test in statistics
- (the t distribution is nearly identical to the normal distribution ($N(0, 1)$), except in very small samples; for your purposes, they are identical)

Implementing the t-test

Step 1: Calculating the test statistic from the data

- Essentially, the t-test asks how far away the sample mean ($\bar{x}$) is from the hypothesized mean (μ_0)
- Distance from the mean is measured in units of standard errors

$$t = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{N}} = \frac{\bar{x} - \mu_0}{se(x)}$$

- t is the test statistic. It measures how many standard errors separate the sample mean and the hypothesized mean.

Testing the hypothesis with a t-test

Step 2: Determine the critical value of the test statistic

- Decide on some level of confidence for our test: How sure do we want to be that we don't mistakenly reject the hypothesis when it is really true?
  - This is the probability of making a Type-I error (rejecting the hypothesis when it is really true). We often use the notation α for this probability.
  - Normally choose $\alpha = 0.05$ (5%).
- Look up **critical value** of test statistic. This is the value that we will reject null hypothesis.
- For $\alpha = 5\%$, the critical value is about 2 (95% of normal distribution is within 2 standard deviations from the mean).
- We can look up the critical value using:
  - using the normal distribution: `qnorm(0.975,0,1)`
  - or using the *t* distribution: `qt(0.975,N-1)`
  - these will be exactly the same when *N* is large, and close to identical even when *N* is small.

Testing the hypothesis with a t-test

Step 3: Compare test statistic to critical value

- We now have a test statistic and a critical value.
- The final step is to compare these two. If the test statistic is bigger than the critical value, we reject the null hypothesis.
  - Test statistic: How many standard errors separate the sample mean and the null hypothesis?
  - Critical value: How many standard errors away should lead us to reject the null hypothesis?

t-test example: coin flipping

We flip a coin 100 times and heads comes up 38 times. Is it a fair coin?

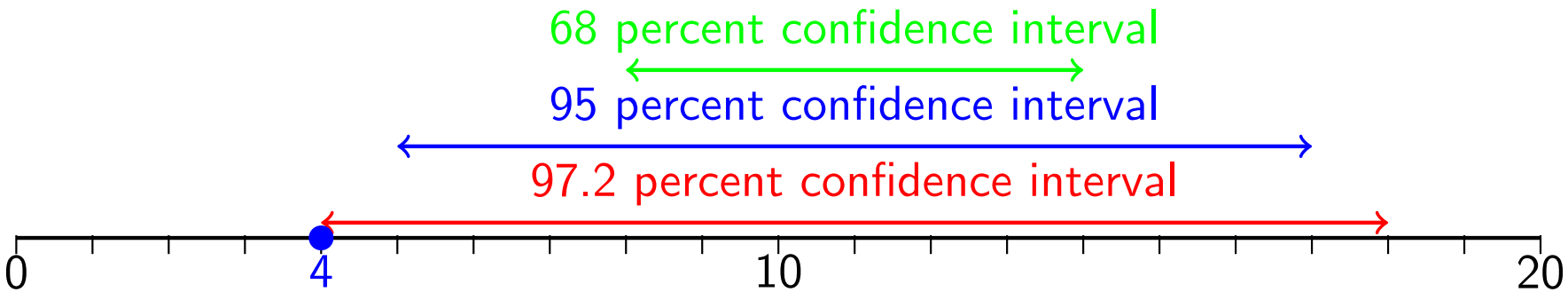
1. Hypothesis: the coin is weighted. Null hypothesis: the coin is no different than a fair coin (i.e., not weighted). $\mu_0 = 0.5$
   - We can generate the data in R:
   - `heads <- c(rep(1,38),rep(0,62))`
2. Test statistic:
   - Mean heads: 0.38 (`mean(heads)`)
   - Standard deviation heads: 0.49 (`sd(heads)`))
   - Standard error $= \frac{0.49}{\sqrt{100}} = 0.05$
   - t-statistic:

$$t = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{N}} = \frac{\bar{x} - \mu_0}{se(x)} = \frac{0.38 - 0.5}{0.05} = 2.4$$

3. Critical value of test statistic:
   - Choose α. Normally $\alpha = 0.05$
   - Find critical value
     - From memory: we know 95% of normal distribution within 2 standard deviations of mean, so $t^{critical} = 2$
     - From R: `qt(0.975,99)` $=$ `qnorm(0.975)` ≈ 1.96
4. Compare: test statistic (2.4) is bigger than critical value (2), so we reject null hypothesis.

p-values

- A p-value expresses how likely the hypothesis is, given the observed data
- Example:
  - Null hypothesis: H_0: true mean is $\mu_0 = 4$
  - Sample mean: $\bar{x} = 11$
  - Sample standard deviation: $\sigma = 31.5$
  - Number of observations in sample: $N = 100$
  - Standard error: $se(x) =?$; $t =?$
  - The hypothesis is more than 2 times the standard error away from the sample mean. This means that if the hypothesis was true, we would only expect to observe a sample mean of 11 very rarely. This is good evidence to reject the hypothesis.
  - We can state the probability (using R): `pt(t,N-1)` $= 2.8\%$



Implementing a t-test in R

- Look up critical value: `qt(1-`$\alpha/2$`,`N`-1)`
- Look up probability given t statistic: `pt(`t`,`N`-1)`
- Univariate t-test:
  - Conduct a t-test that true mean is zero: `t.test(data_vector)`
  - Conduct a t-test that true mean is zero: `lm(data_vector ~ 1)`
- Bivariate t-test:
  - Compare two groups: `t.test(outcome_variable ~ group_variable)`
  - Compare two groups: `lm(outcome_variable ~ group_variable)`