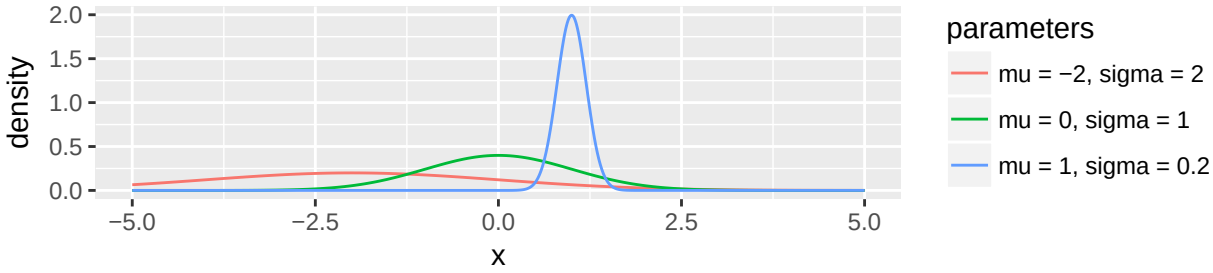
Z-Scores and t-tests

Ideas from Chapters 5, 17, and 20

Reminder of Background

Normal Model (Chapter 5)

- $Y \sim \text{Normal}(\mu, \sigma)$
- Read: "Y is modeled as following a normal distribution with mean μ and standard deviation σ"
- Here, Y is the value of a **quantitative** variable associated with one randomly sampled item from the population
- μ and σ are **parameters** that control the center (mean) and spread (standard deviation) of the distribution



Central Limit Theorem (Chapter 17)

- $Y_1, Y_2, \ldots, Y_n$ are **independent** observations of a **quantitative variable**
- Population has mean μ and standard deviation σ
- For large enough n, the sampling distribution of the sample mean $\bar{Y}$ is approximately:
  - $\bar{Y} \sim \text{Normal}(\mu, \sigma/\sqrt{n})$
- This describes the disitribution of values for the sample mean from all possible samples of size n from the population.

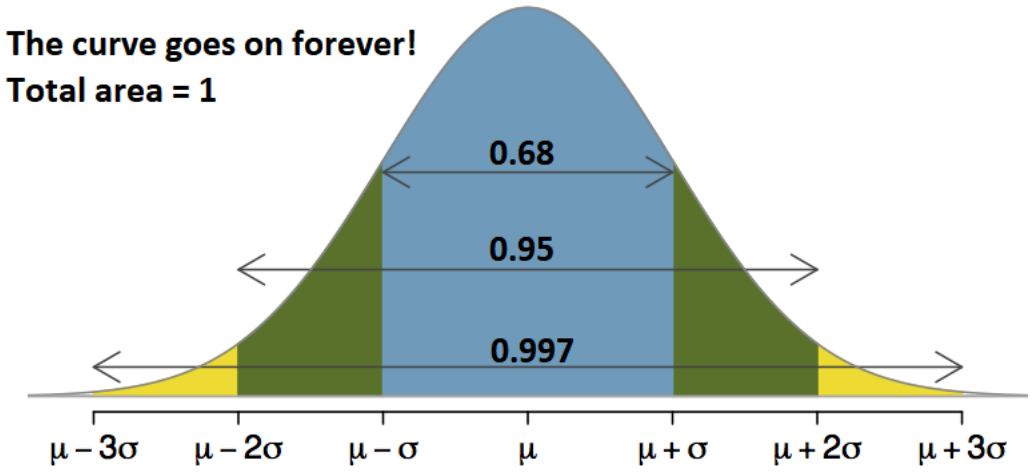
Example 1

Suppose that in 20 MPH speed zones, Americans drive 24 MPH on average, with a standard deviation of 4 MPH. We take a random sample of 100 drivers in a 20 MPH speed zone, and calculate the mean speed of drivers in this sample. What is the sampling distribution of the sample mean speed?

Probabilities for the Normal Model, Z-scores (see Chapter 5)

Suppose $Y \sim \text{Normal}(\mu, \sigma)$

- 68% of the time, the value of Y will be within 1 standard deviation of the mean
- 95% of the time, the value of Y will be within 2 standard deviations of the mean
- 99.7% of the time, the value of Y will be within 3 standard deviations of the mean



Key Idea: To calculate probabilities involving the normal distribution, we need to find out how many standard deviations away from the mean a particular number is. This is what a z-score tells us!

If $X \sim \text{Normal}(\mu, \sigma)$, the z-score of a number x is:

$z = \frac{x-\mu}{\sigma}$

Note: If $X \sim \text{Normal}(\mu, \sigma)$, then $Z \sim \text{Normal}(0, 1)$

Example 2

Suppose that in 20 MPH speed zones, Americans drive 24 MPH on average, with a standard deviation of 4 MPH.

(a) We randomly select a single driver and measure their speed. What is the probability that they are driving between 16 and 32 MPH?

(b) We randomly select a single driver and measure their speed. What is the probability that they are driving more than 32 MPH?

(c) We randomly select 100 drivers and measure their speeds. What is the probability that the mean speed of these drivers is more than 20.4 MPH?

Example 3

Among the population of all babies born in December of 1998, the standard deviation of the gestation time was about 2.6 weeks. Suppose I take a sample of 100 babies from this population.

If the population mean gestation time was 40 weeks, what would be the probability that the sample mean would be greater than 40.52 weeks?

Hypothesis Tests About a Population Mean with the Normal Distribution

Suppose we know the population standard deviation, σ, and we want to test a hypothesis about whether a population mean is equal to a number μ_0. (For example, if I'm doing a test of whether the population mean is equal to 12, then $\mu_0 = 12$). We take a sample of size n and calculate the sample mean, $\bar{x}$.

We need to calculate the **p-value**: the probability of obtaining a test statistic at least as extreme as what we observed based on our sample, assuming the null hypothesis is true.

What are the null and alternative hypotheses for the test? (Express them in terms of μ_0)

If the null hypothesis is true, what is the sampling distribution of the sample mean, $\bar{X}$? (Express it in terms of μ_0, σ, and n)

Our test statistic will be the z-score of the sample mean. Write down a formula for how you would calculate this. (Express it in terms of μ_0, σ, and n)

If the null hypothesis is true, what is the sampling distribution of the test statistic?

Suppose you take a sample, calculate the sample mean, and then the z-score of the sample mean, and you get a test statistic of 2. Draw a picture of the sampling distribution of the test statistic, and shade in the region corresponding to the p-value.

Example 4

Among the population of all babies born in December of 1998, the standard deviation of the gestation time was about 2.6 weeks. Suppose I take a sample of 100 babies from this population and obtain a sample mean of 39.22 weeks. Conduct a hypothesis test of the claim that the mean gestation time is 40 weeks, at the $\alpha = 0.05$ significance level.

(For this example, let's skip all the assumption checks)

Hypothesis Tests About a Sample Mean with the t Distribution (Final Answer)

Problem:

We usually don't know the population standard deviation σ, so we can't actually calculate $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$

Solution:

Estimate the population standard deviation with the sample standard deviation.

New test statistic:

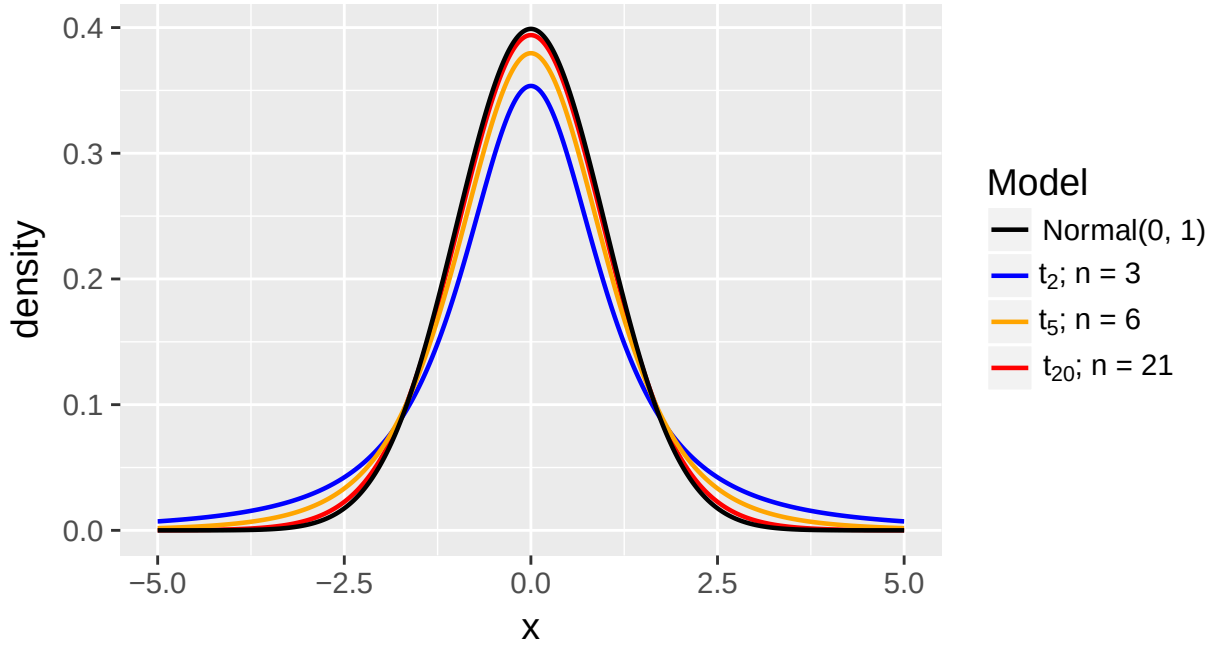
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

,

where $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$ is the standard deviation of the data in the sample.

Facts:

- $t \sim t_{n-1}$
- Read as "t follows a t distribution with $n-1$ degrees of freedom"
- The degrees of freedom of $n-1$ matches the denominator of $n-1$ in the sample standard deviation
- The t distribution is similar to the Normal$(0,1)$, but t has more probability in the tails
- As the degrees of freedom increases, the t becomes more like a Normal$(0,1)$



Example 5

Suppose I take a sample of 100 babies who were born in December 1998 and record their gestation times. Conduct a hypothesis test of the claim that the mean gestation time is 40 weeks, at the $\alpha = 0.05$ significance level. As part of your work, show how the p-value is calculated. No need to check the conditions for inference for this example. You may use the following R output:

```
babies_sample %>%
  summarize(
    mean_gestation = mean(gestation),
    sd_gestation = sd(gestation)
  )
```

```
## # A tibble: 1 x 2
##   mean_gestation sd_gestation
##            <dbl>        <dbl>
## 1          38.83     2.238574
```

```
t.test(~gestation, mu = 40, data = babies_sample)
```

```
## ~gestation
```

```
##
##  One Sample t-test
##
## data:  gestation
## t = -5.2265, df = 99, p-value = 9.611e-07
## alternative hypothesis: true mean is not equal to 40
## 95 percent confidence interval:
##  38.38582 39.27418
## sample estimates:
## mean of x
##     38.83
```

```
pt(-5.226542, df = 99)
```

```
## [1] 4.805429e-07
```

```
2 * pt(-5.226542, df = 99)
```

```
## [1] 9.610858e-07
```

Example 6 (also see Lab 15 on Gryd; adapted from SDM4 20.44)

Consumer Reports tested 11 brands of vanilla yogurt and found the following numbers of calories per serving:

```
yogurt
```

```
##    calories
## 1       130
## 2        60
## 3       150
## 4       120
## 5       120
## 6       110
## 7       170
## 8       160
## 9       110
## 10      130
## 11       90
```

(a) Check the conditions for inference.

(b) A nutrition guide claims that you will get an average of 120 calories from a serving of vanilla yogurt. Conduct an appropriate hypothesis test and state your conclusion. As part of the hypothesis test, find the p-value for the test "manually". When you're doing this, draw a picture of a t distribution, show where the test statistic is, and shade in the region corresponding to the p-value.

Note: Lab 15 on Gryd sets up data and a place where you can run the neccessary R commands.