Day 27 - Single mean (parametric)

Learning outcomes

In this section, we explore how to formally make inference about a single mean using parametric (theory-based) methods.

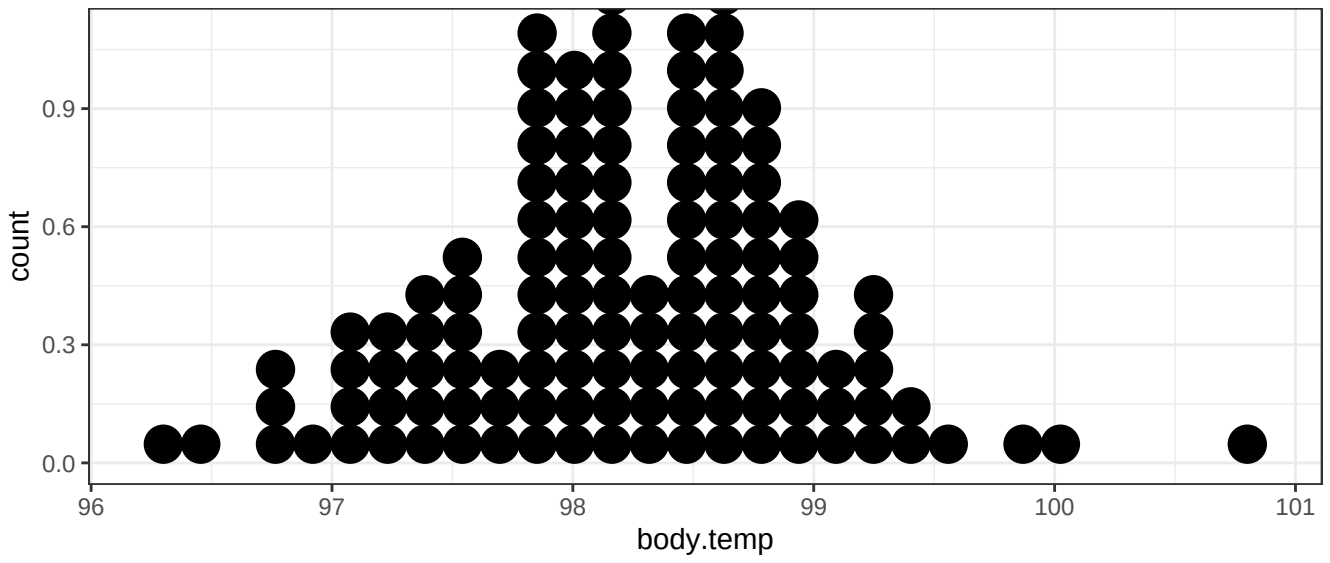
Motivating example

Body temperatures It is commonly stated that the average human body temperature is 98.6 degrees Fahrenheit. Do data support this claim? Body temperature readings were collected on 130 volunteer adults. Is there evidence that the true mean human body temperature differs from 98.6 degrees Fahrenheit?

```
# load data
temps <- read_csv("thermometry.csv")
head(temps, n = 2)
```

```
# A tibble: 2 x 2
  body.temp heart.rate
      <dbl>      <dbl>
1      96.3         70
2      96.7         71
```

```
# create dot plot
temps %>% ggplot() +
  geom_dotplot(aes(x = body.temp), method = "histodot") + theme_bw() + ylim(c(0, 1.1))
```




```
# summary statistics
temps %>% summarize(mean = mean(body.temp), sd = sd(body.temp), n = n())
```

```
# A tibble: 1 x 3
   mean    sd     n
  <dbl> <dbl> <int>
1  98.2 0.733   130
```

Review

What are the observational units for this study?

What variables are collected? Are they categorical or quantitative?

What is the research question?

To whom can we generalize the results of this inference?

More review

What is the parameter of interest?

What is the sample statistic?

What are the hypotheses for this test?

Conditions for inference

Conditions

- **Independence**: We assume that observations are _____.
- **Central Limit Theorem**: We require that sample size be sufficiently large for the CLT to apply. For a single quantitative variable, the sample size required depends upon the _____ of the _____.
  - If the population distribution is approximately _____ with no outliers, very few observations are required.
  - If the population distribution is moderately _____ with some outliers, more observations are required.
  - If the population distribution is heavily _____ with outliers, many observations are required.
  - The bootstrap sampling distribution of the sample mean can be used to help understand whether the CLT conditions are met.

Distribution of the statistic

If the CLT conditions are met, the distribution of the sample mean is normal with mean μ and standard deviation

$$SD(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

To estimate the standard deviation of the sample mean, we use the _____ of the sample mean

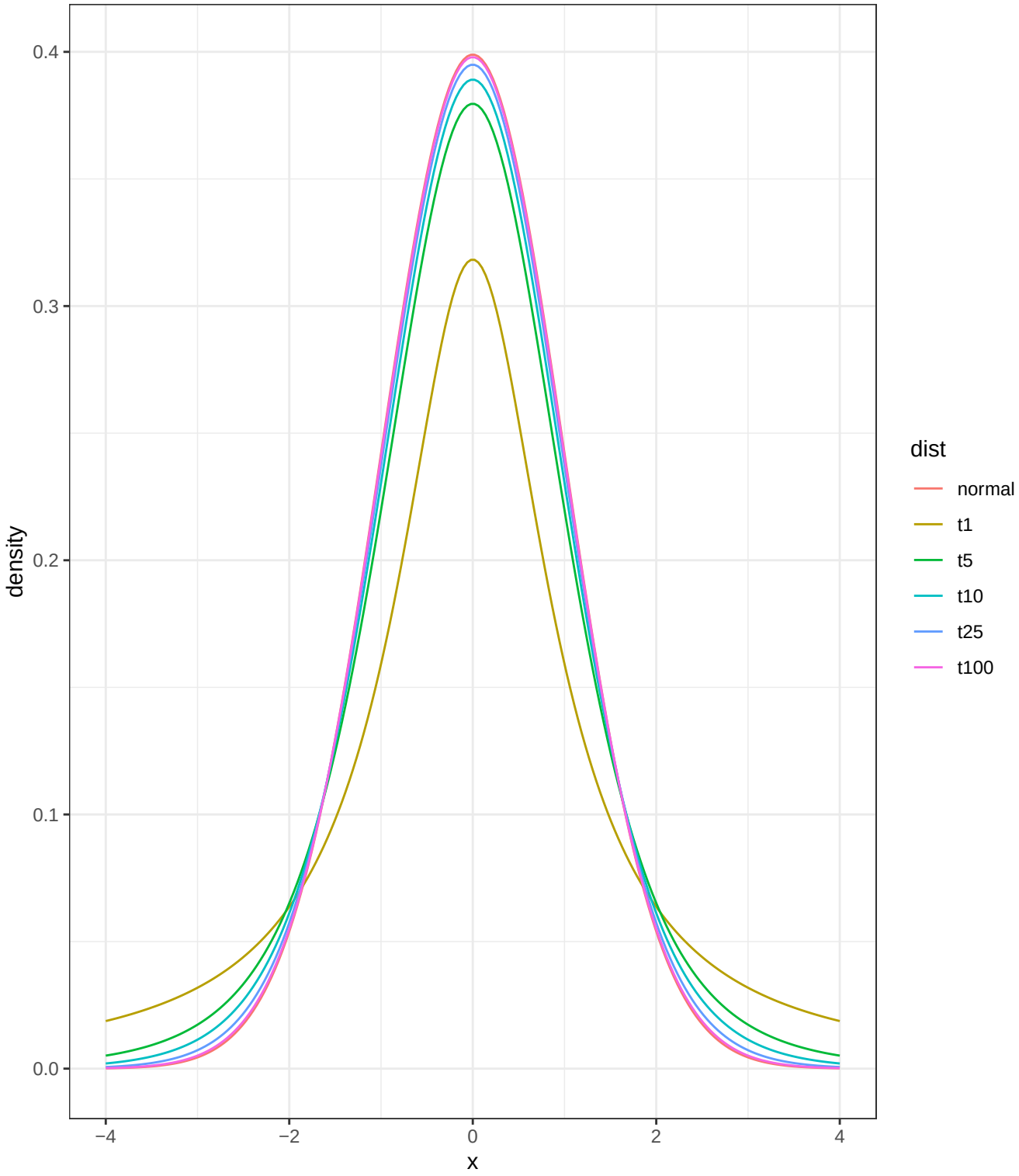
$$SE(\bar{x}) = \frac{s}{\sqrt{n}}$$

Notice that the distribution of the sample mean depends upon _____ unknown parameters: μ and σ. To account for the additional variability that results from estimating two unknown parameters, the _____ is used to make inference about sample means.

The t-distribution

> **Note**
>
> The _____ is similar to the normal distribution, but has more probability in the _____. The standard t-distribution is always centered at _____ and is described by the _____.

```
tibble(
  x = seq(-4, 4, length.out = 200)
) %>%
  mutate(
    normal = dnorm(x),
    t1 = dt(x, df = 1),
    t5 = dt(x, df = 5),
    t10 = dt(x, df = 10),
    t25 = dt(x, df = 25),
    t100 = dt(x, df = 100)
  ) %>%
  pivot_longer(normal:t100, names_to = "dist", values_to = "density") %>%
  mutate(
    dist = factor(dist, levels = c("normal", "t1", "t5", "t10", "t25", "t100"))
  ) %>%
  ggplot() +
  geom_line(aes(x = x, y = density, col = dist)) +
  theme_bw()
```


0.4
0.3
0.2
0.1
0.0
density
−4
−2
0
2
4
x
dist
normal
t1
t5
t10
t25
t100

Hypothesis testing

Hypothesis testing

Under the _____, the sampling distribution of

$$t = \frac{\text{statistic - null}}{SE(\text{statistic})} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

is t with $n - 1$ degrees of freedom if the _____ conditions are met.

Example: Hypothesis test for the body temps example:

1. Assess the conditions for hypothesis testing.

   First, we assume that the participants' body temperature are independent of one another. This is probably reasonable, assuming we didn't include family members or something like that. The sample distribution of body temperatures is fairly symmetric to begin with, and the sample size is quite large ($n = 130$). Therefore, the sample size is likely large enough for the CLT to provide approximately normally distributed sampling distributions of the sample mean.

2. Calculate the observed statistic, $\bar{x}$, the standard deviation, s, and the test statistic, t.

```
(temp_mean <- mean(temps$body.temp))
```

```
[1] 98.24923
```

```
(temp_sd <- sd(temps$body.temp))
```

```
[1] 0.7331832
```

```
(t <- (temp_mean - 98.6) / (temp_sd/sqrt(130)))
```

```
[1] -5.454823
```

3. Calculate the p-value.

```
# add both tails
pt(t, df = 129) + 1 - pt(abs(t), df = 129)
```

```
[1] 2.410632e-07
```

```
# shortcut
2*pt(-abs(t), df = 129)
```

```
[1] 2.410632e-07
```

4. Interpret p-value and state a conclusion.

Confidence intervals

Confidence intervals

Assuming _____ is true has **no affect on the standard error.** A $(1-\alpha)\%$ confidence interval is calculated in the usual way, this time using a t-multiplier:

$$\begin{aligned}\text{obs stat} \pm \text{margin of error} &= \bar{x} \pm t^*_{1-\alpha/2,df} SE(\bar{x}) \\ &= \bar{x} \pm t^*_{1-\alpha/2,df} \frac{s}{\sqrt{n}}\end{aligned}$$

where $t^*_{1-\alpha/2,df}$ is the $1-\alpha/2$ quantile of the t-distribution with df degrees of freedom.

5. Compute a 99% confidence interval for $\pi_1 - \pi_2$.

```
temp_mean + c(-1,1)*qt(.995, 129)*(temp_sd/sqrt(130))
```

```
[1] 98.08111 98.41735
```

6. Interpret the 99% confidence interval.