

# Research Design, Fall 2021

**04: a bit more probability review; sampling; statistical tests**

general feedback for research reviews

research reviews, general guidance

Title: homework 1

This paper described A and ...

research reviews, general guidance

~~Title: homework 1~~

Why should I be reading this document? What is the main point?

The authors of ~~this paper~~ described A and ...

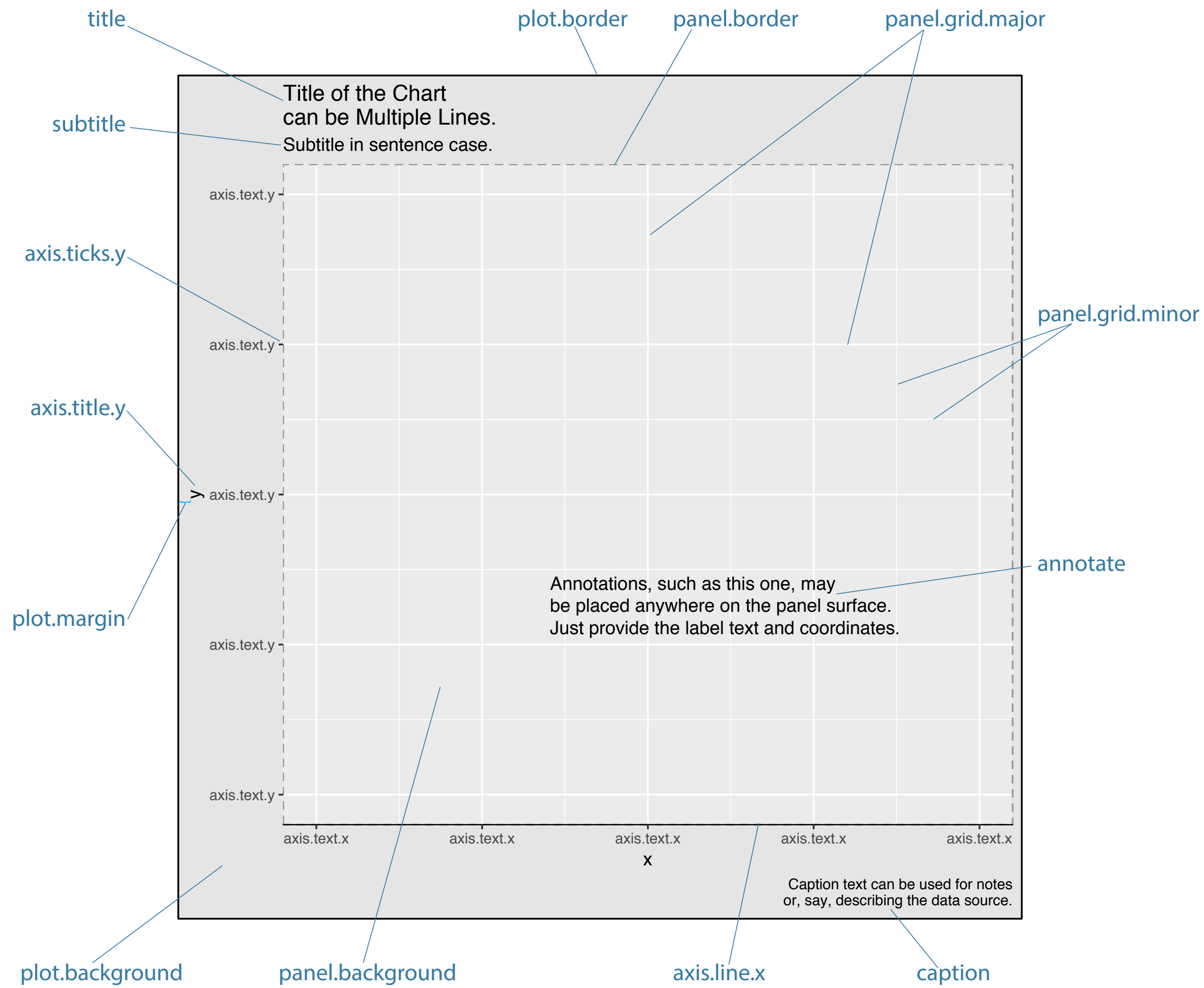
To what paper are you referring? Always introduce new ideas before using language that refers back. And properly cite your sources.

**Also review:**

Harris, Joseph. *Rewriting: How to Do Things with Texts*.  
Second edition. Logan: Utah State University Press, 2017.

tips on graphics as a tool for exploring and communicating research design

# An aside on graphics to communicate research designs,



# Coding graphic elements, *example in R/GGplot2*

```
# load grammar of graphics
library(ggplot2)
```

```
p <-
```

```
# functions for data ink
```

```
ggplot(data = <data>,
       mapping = aes(<aesthetic> = <variable>,
                     <aesthetic> = <variable>,
                     <...> = <...>)) +
  geom_<type>(<...>) +
  scale_<mapping>_<type>(<...>) +
  coord_<type>(<...>) +
  facet_<type>(<...>) +
  <...> +
```

```
# functions for non-data ink
```

```
labs(<...>) +
theme(<...> = <...>) +
annotate(<...>) +
<...>
```

```
element_blank()
element_line(<...> = <...>)
element_rect(<...> = <...>)
element_text(<...> = <...>)
```

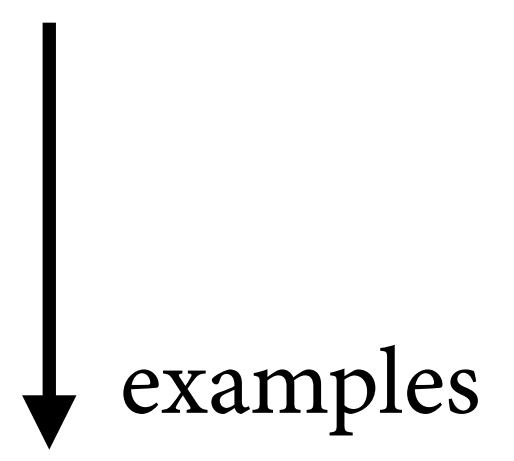
# Drawing functions,

```
stat_function(
  fun = <function name>,
  args = list(parameter1 = <...>, parameter2 = <...>),
  geom = "<type>",
  ...)
```

a bit more statistics and probability

R's probability functions, **p**robability density (PDF), cumulative **d**istribution (CDF), **q**uantile, **r**andom generation

<b>p</b> <probability function name>	<b>p</b> robability density function
<b>d</b> <probability function name>	cumulative <b>d</b> istribution function
<b>q</b> <probability function name>	<b>q</b> uantile function
<b>r</b> <probability function name>	<b>r</b> andom generation function



---

<i>normal</i>	<i>negative binomial</i>
<i>student t</i>	<i>gamma</i>
<i>bernoulli</i>	<i>cauchy</i>
<i>binomial</i>	<i>100s more</i>
<i>poisson</i>	



law of large numbers, as  $n \rightarrow \infty$ ,  $\bar{x} \rightarrow \mu$

```
set.seed(29914)

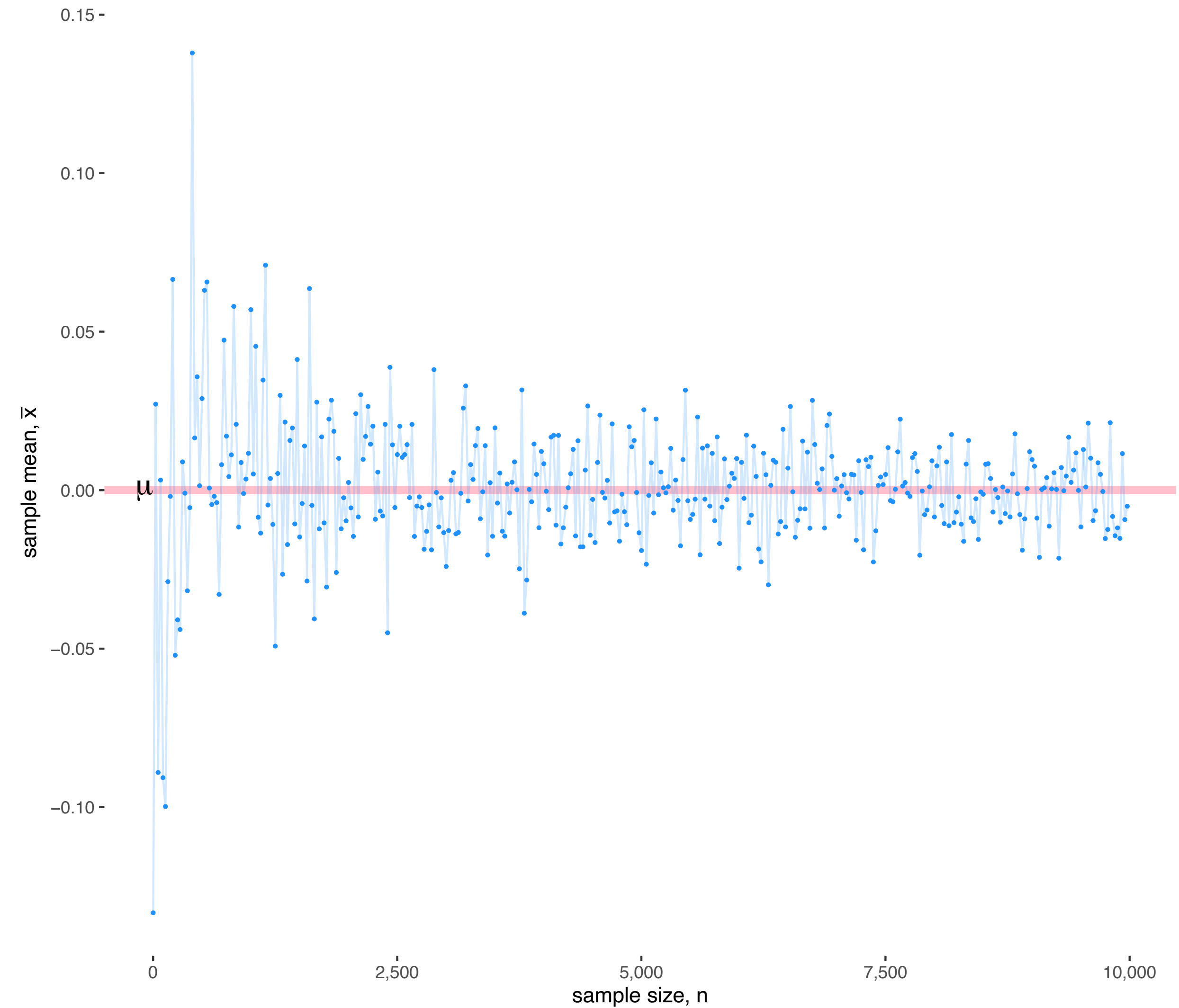
n      <- seq(1, 10000, by = 25)
mu     <- 0
sigma  <- 1
x_bar  <- sapply(n, FUN = function(n) {mean( rnorm(n, mu, sigma) )} )
```

```
library(ggplot2); library(ggthemes); library(latex2exp)
theme_set( theme_tufte(base_family = "sans") )

ggplot() +
  scale_x_continuous(labels = scales::comma) +
  geom_hline(yintercept = mu, color = "pink", lwd = 2) +
  geom_point(aes(n, x_bar), size = 0.5, color = "dodgerblue") +
  geom_line(aes(n, x_bar), alpha = 0.2, color = "dodgerblue") +
  annotate("text", 0, mu, hjust = 1, size = 16/.pt, label = TeX("$\\mu$")) +
  labs(x = "sample size, n", y = TeX("sample mean, $\\bar{x}$"))
```

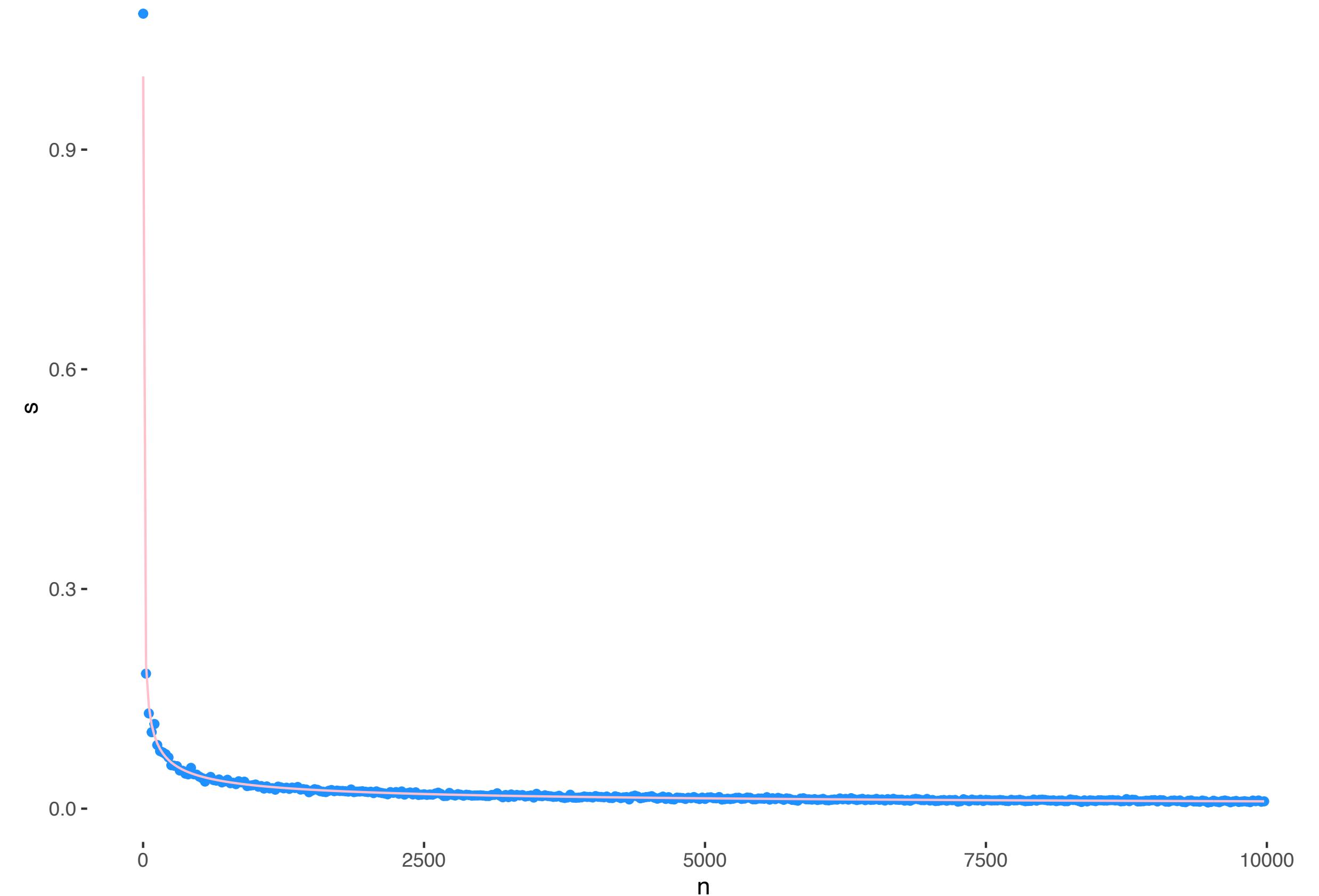
Experiment with different values of  $n$ ,  $\mu$ , and  $\sigma$  to see how  $\bar{x}$  compares with  $\mu$ .

Experiment with different seed values and compare results. How would you describe the uncertainty of  $\bar{x}$ ?



central limit theorem, standard deviation  $s$  of  $\bar{x} = \frac{\sigma}{\sqrt{n}}$

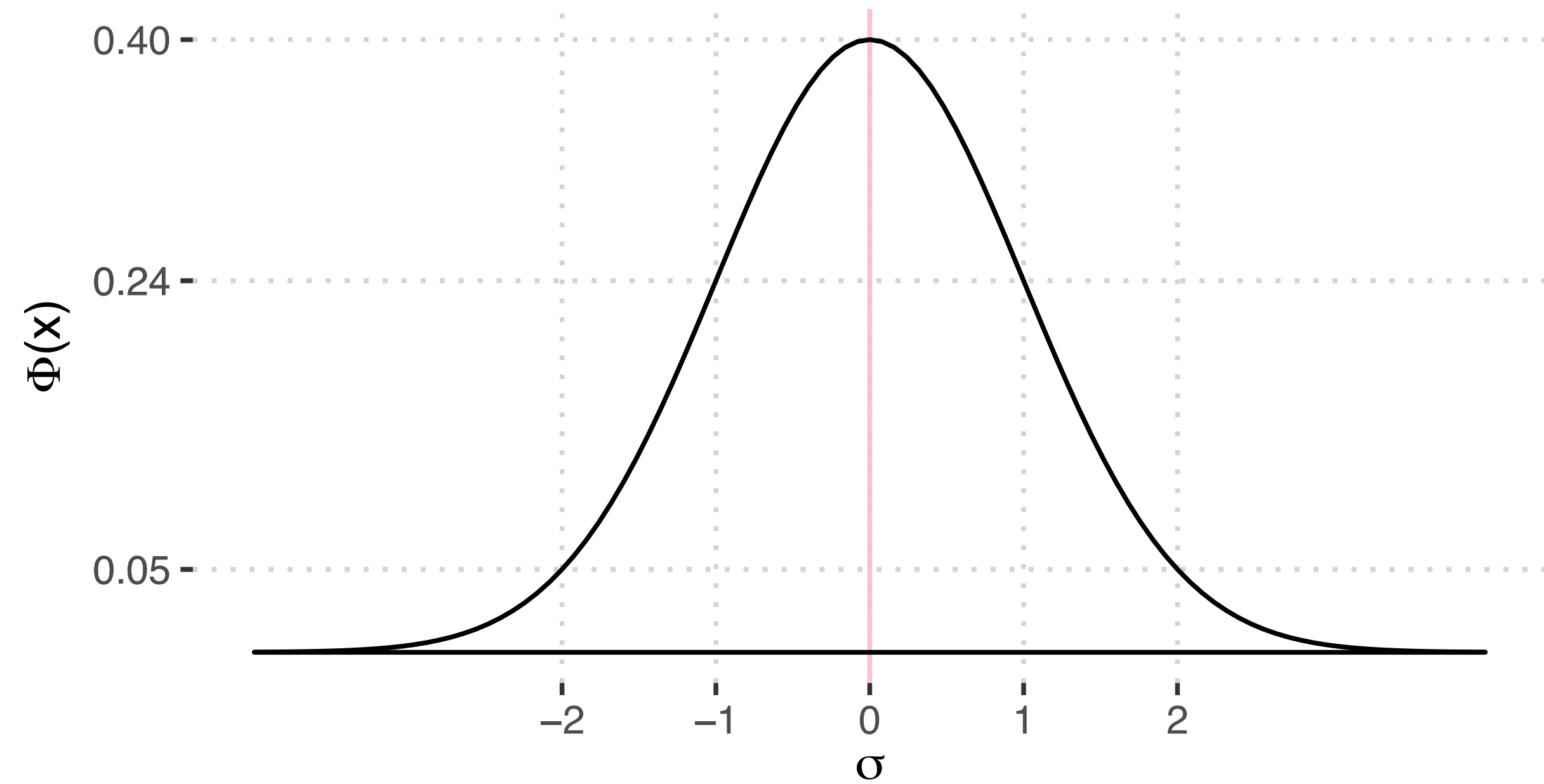
```
sample_mean <- function() {  
  sapply(n, FUN = function(n) {mean( rnorm(n, mu, sigma) )} )  
}  
  
x_bar <- replicate(1000, sample_mean() )  
s <- apply(x_bar, 1, sd)  
  
ggplot() +  
  geom_point(aes(n, s), color = "dodgerblue") +  
  geom_line(aes(n, sigma / sqrt(n) ), color = "pink")
```



Experiment with different values of  $\sigma$  and  $\mu$  and replications to check the relationships empirically.

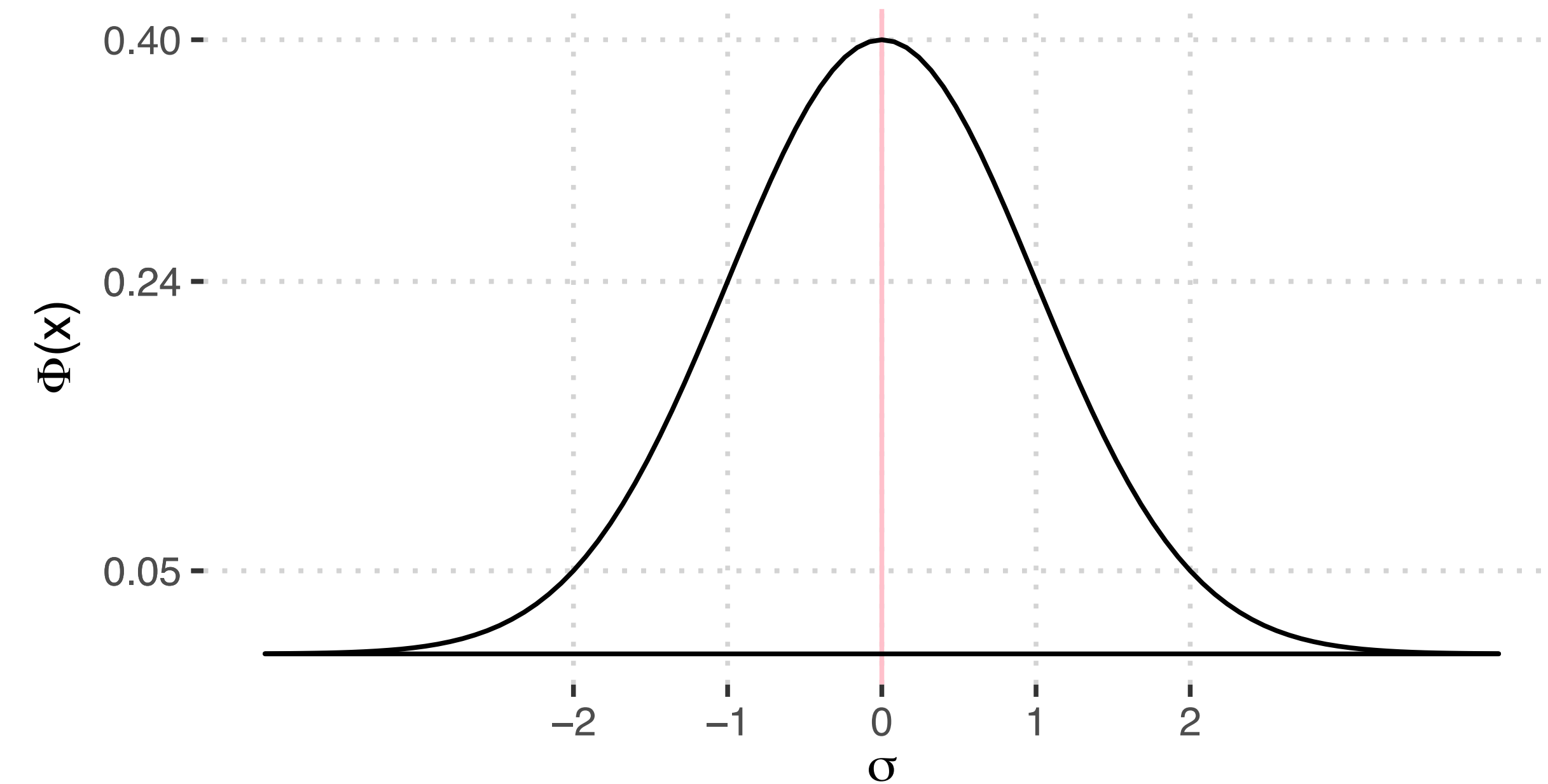
# probability, probability density functions v. continuous distribution functions — e.g., the standard normal $\Phi$

```
pdf <- ggplot() +  
  theme(panel.grid.major = element_line(color = "lightgray", linetype = "dotted")) +  
  scale_x_continuous(breaks = seq(-2, 2)) +  
  scale_y_continuous(breaks = dnorm(seq(-2, 2)), labels = scales::comma) +  
  geom_vline(xintercept = 0, color = "pink") +  
  stat_function(fun = dnorm,  
    args = list(mean = 0, sd = 1),  
    geom = "density",  
    xlim = c(-4, 4)) +  
  labs(x = TeX("$\\sigma$"), y = TeX("$\\Phi(x)$"))
```

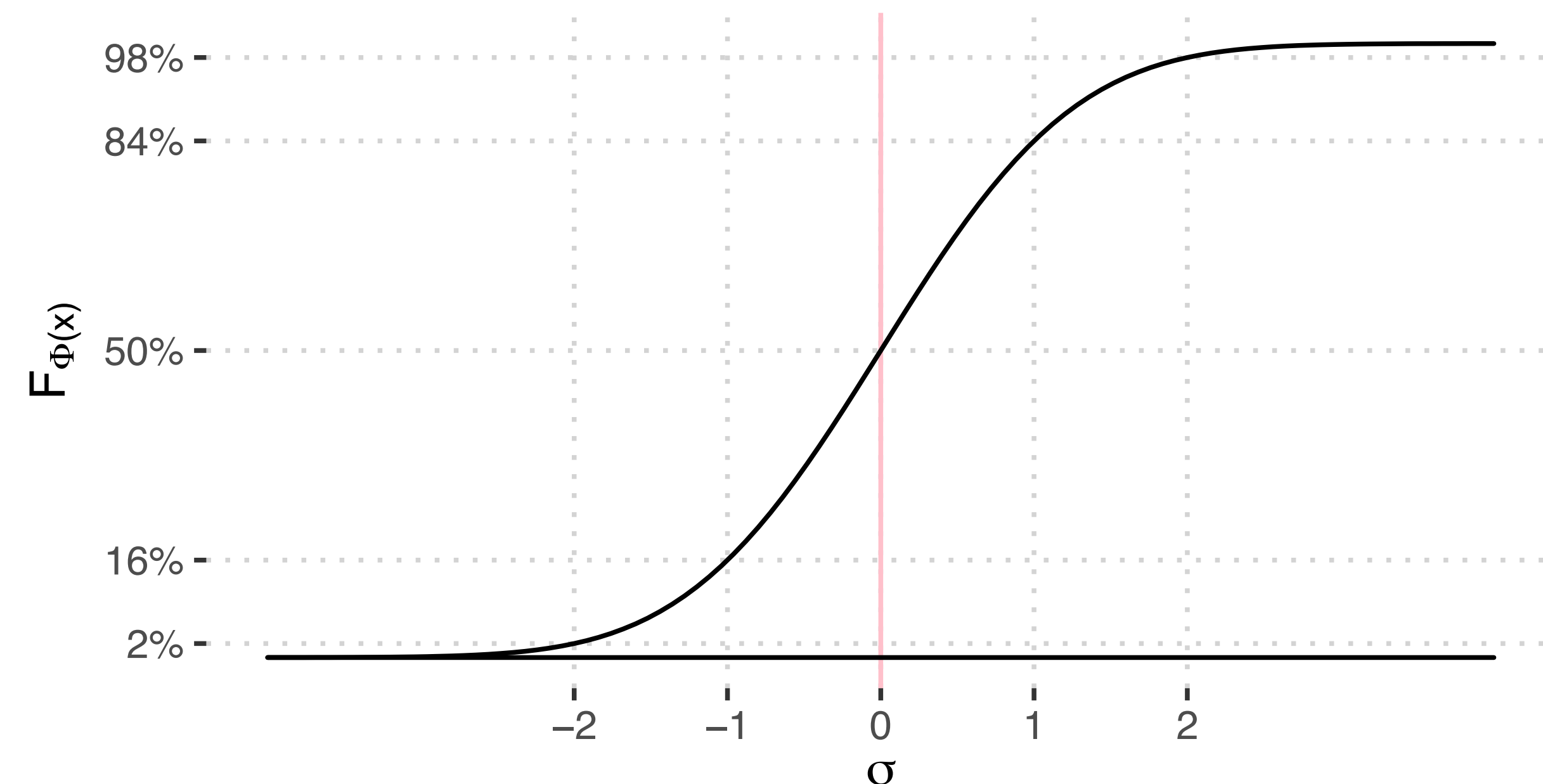


# probability, probability density functions v. continuous distribution functions — e.g., the standard normal $\Phi$

```
pdf <- ggplot() +  
  theme(panel.grid.major = element_line(color = "lightgray", linetype = "dotted")) +  
  scale_x_continuous(breaks = seq(-2, 2)) +  
  scale_y_continuous(breaks = dnorm(seq(-2, 2)), labels = scales::comma) +  
  geom_vline(xintercept = 0, color = "pink") +  
  stat_function(fun = dnorm,  
    args = list(mean = 0, sd = 1),  
    geom = "density",  
    xlim = c(-4, 4)) +  
  labs(x = TeX("$\\sigma$"), y = TeX("$\\Phi(x)$"))
```



```
cdf <- ggplot() +  
  theme(panel.grid.major = element_line(color = "lightgray", linetype = "dotted")) +  
  scale_x_continuous(breaks = seq(-2, 2)) +  
  scale_y_continuous(breaks = pnorm(seq(-2, 2)),  
    labels = scales::label_percent(accuracy = 1)) +  
  geom_vline(xintercept = 0, color = "pink") +  
  stat_function(fun = pnorm,  
    args = list(mean = 0, sd = 1),  
    geom = "density",  
    xlim = c(-4, 4)) +  
  labs(x = TeX("$\\sigma$"), y = TeX("$F_{\\Phi(x)}$"))
```



more on sampling

sampling, a few of many approaches

simple random

convenience

interval

cluster

quota

stratified

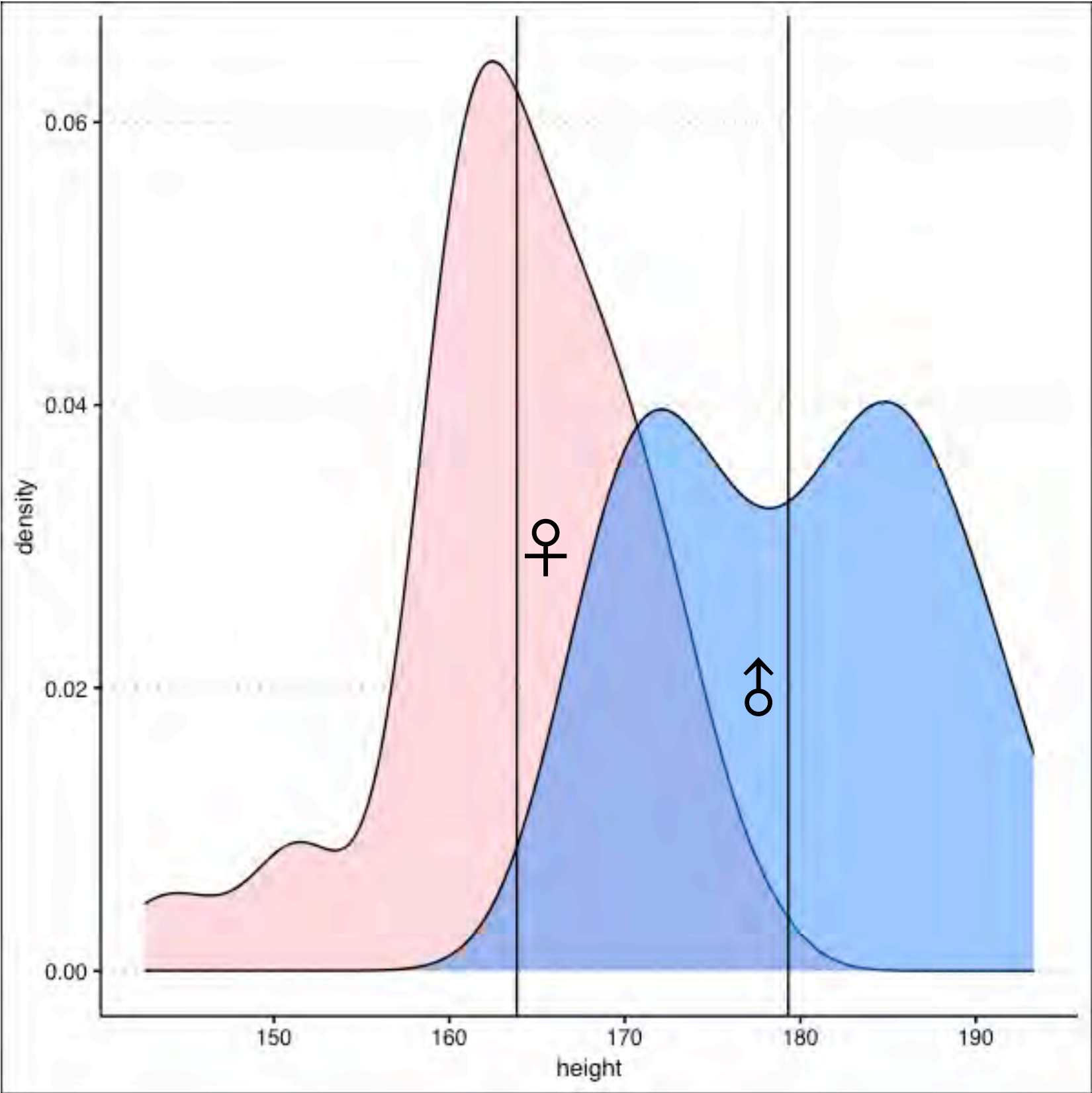
quasi-experimental

statistical tests

Example, recall our random sample of heights in NYC

Sample mean for both groups,

male	x_bar
-----	-----
FALSE	163.8453
TRUE	179.3198



How do we decide whether to reject  $H_0$  in favor of  $H_A$ ?

$$H_0 : \overline{\text{height}}_{\text{men}} = \overline{\text{height}}_{\text{women}}$$

$$H_A : \overline{\text{height}}_{\text{men}} \neq \overline{\text{height}}_{\text{women}}$$

We need some kind of test!



# A general procedure for a statistical test

Assume an appropriate *probability model* to describe the behavior of the random variable under investigation.

Define a *null* hypothesis and an *alternative* hypothesis that permits meaningful conclusions.

Specify a *test statistic*.

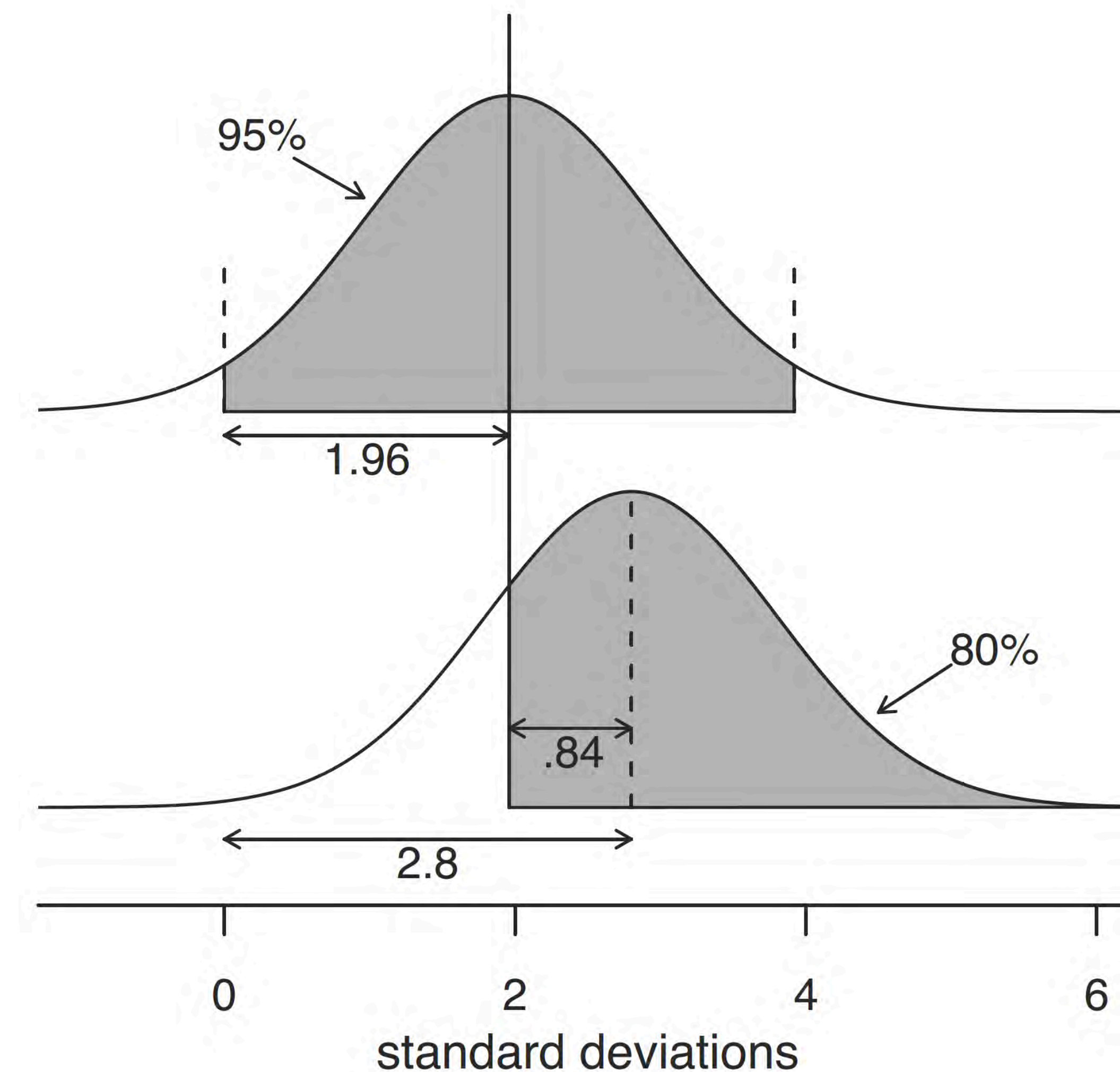
Choose a *level of significance*  $\alpha$  for the test.

Determine a *distribution* and *critical region* of the test statistic.

Calculate a *value* of the test statistic from a random sample of data.

Accept or reject  $H_0$  by comparing the calculated value of the test statistic with the values defining the critical region. 🤖

Generic normal distribution, distance to zero, at a 95 percent *confidence interval* and 80 percent *power*



statistical tests, comparing sample mean to normal distribution with known  $\mu$  and  $\sigma$  — *z-statistic*

$$H_0 : \mu = 2.7, H_A : \mu < 2.7$$

```
set.seed(92)

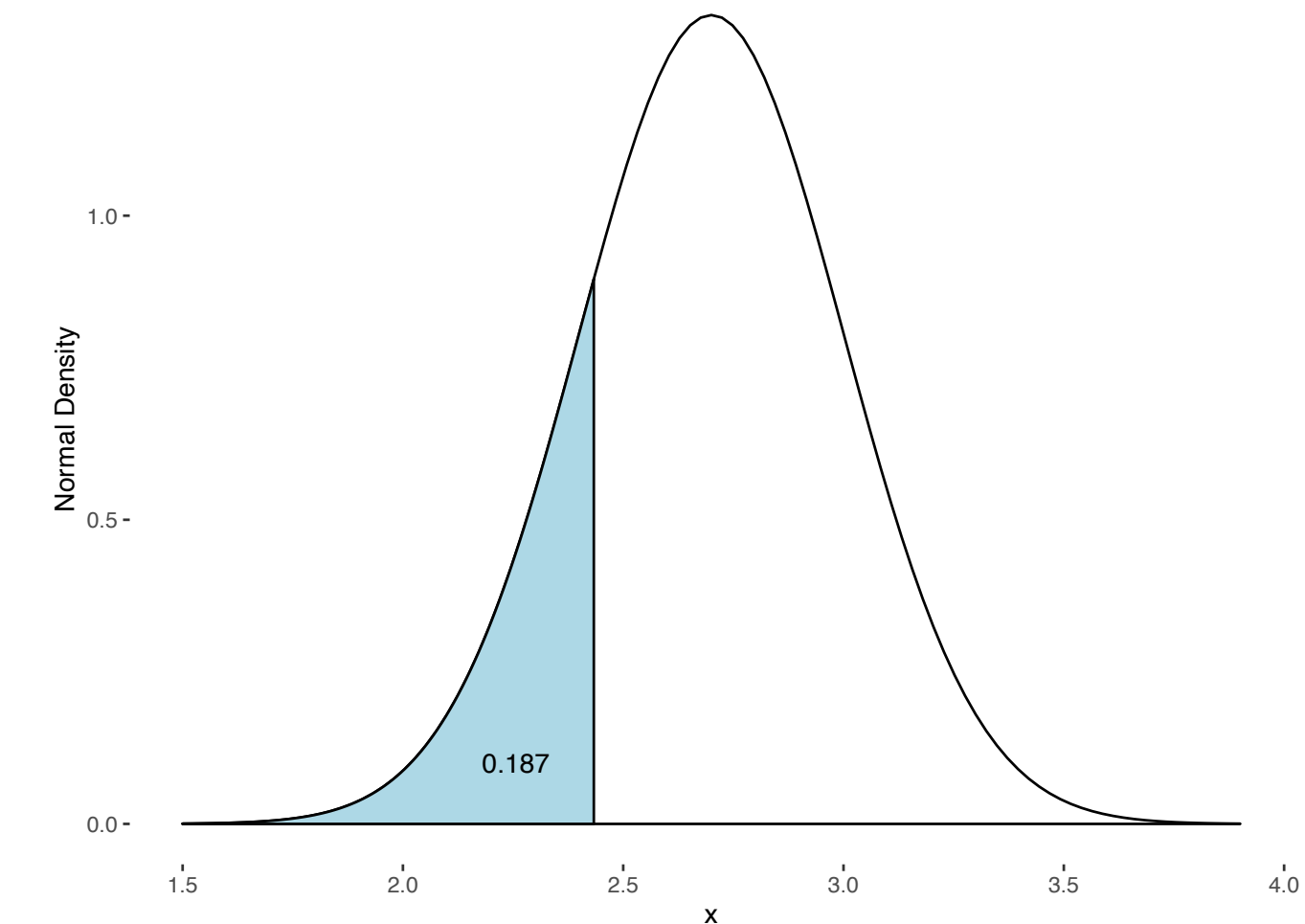
n <- 30
mu <- lambda <- 2.7
sigma <- sqrt(lambda)
x <- rpois(n, lambda)
```

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}, p = F_{\Phi}(z)$$

```
x_bar <- mean(x)

z <- ( x_bar - mu ) / ( sigma / sqrt(n) )
p <- pnorm(z)
```

Given the null hypothesis, about  
19 percent of experiments have  $\bar{x} \leq 2.43$



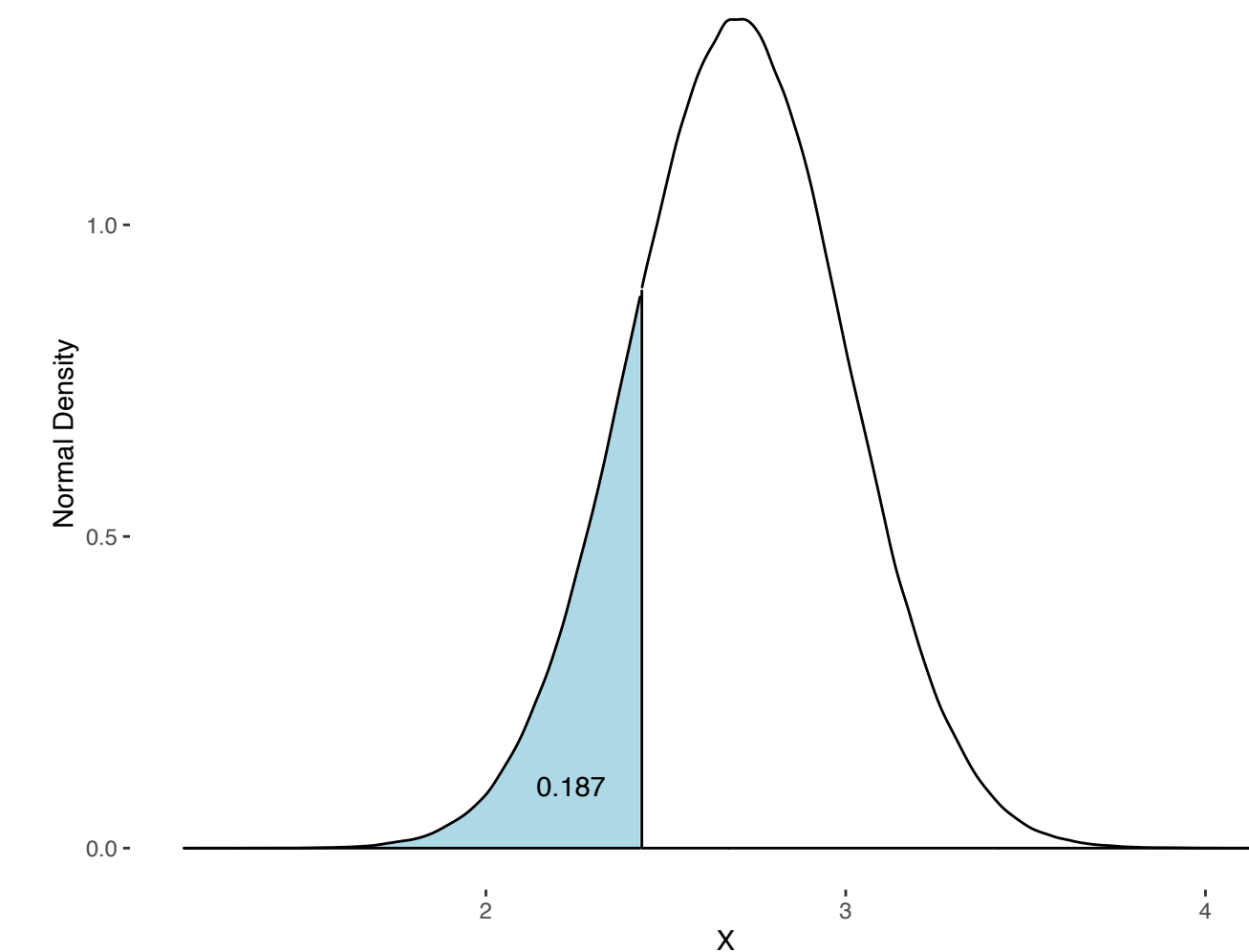
```
ggplot() +
  stat_function(fun = dnorm,
               args = list(mean = mu, sd = sigma / sqrt(n)),
               geom = "density",
               fill = "white",
               xlim = c(mu - 4 * sigma / sqrt(n),
                       mu + 4 * sigma / sqrt(n)) ) +
  stat_function(fun = dnorm,
               args = list(mean = mu, sd = sigma / sqrt(n)),
               geom = "density",
               fill = "lightblue",
               xlim = c(mu - 4 * sigma / sqrt(n), x_bar)) +
  annotate("segment", x = x_bar, xend = x_bar,
          y = 0, yend = dnorm(x_bar, mu, sigma / sqrt(n))) +
  annotate("text", x = x_bar - 0.1, y = 0.1, hjust = 1,
          label = format(p, digits = 3)) +
  scale_x_continuous(breaks = seq(-1.5, 4, by = 0.5)) +
  labs(y = "Normal Density")
```

# statistical tests, comparing sample mean to normal distribution *with simulation*

$$H_0 : \mu = 2.7, H_A : \mu < 2.7$$

```
X <- rnorm(1e6, mu, sigma / sqrt(n) )  
d <- data.frame( density(X)[1:2] )
```

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}, p = F_{\Phi}(z)$$



```
ggplot(d) +  
  theme(legend.position = "") +  
  geom_ribbon(aes(x = x,  
                 ymin = 0,  
                 ymax = y,  
                 fill = x < x_bar),  
            color = "black") +  
  scale_x_continuous(breaks = seq(-1.5, 4, by = 0.5)) +  
  scale_fill_manual(values = c("white", "lightblue")) +  
  annotate("segment", x = x_bar, xend = x_bar,  
          y = 0, yend = dnorm(x_bar, mu, sigma / sqrt(n))) +  
  annotate("text",  
          x = x_bar - 0.1, y = 0.1, hjust = 1,  
          label = format( mean( X < x_bar ), digits = 3 ) ) +  
  labs(x = "X", y = "Normal Density")
```

**Simulation gives us the same answer:**  
given the null hypothesis, about 19 percent of  
experiments have  $\bar{x} \leq 2.43$

statistical tests, if **unknown**  $\sigma$ , can use sample standard deviation  $s$  and student's t distribution — *t-statistic*

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}, p = F_{\Phi}(z)$$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}, p = F_T(t)$$

Of note: as  $s \rightarrow \sigma \mid n \rightarrow \infty, t \rightarrow z$  and student's t distribution converges towards the normal distribution.

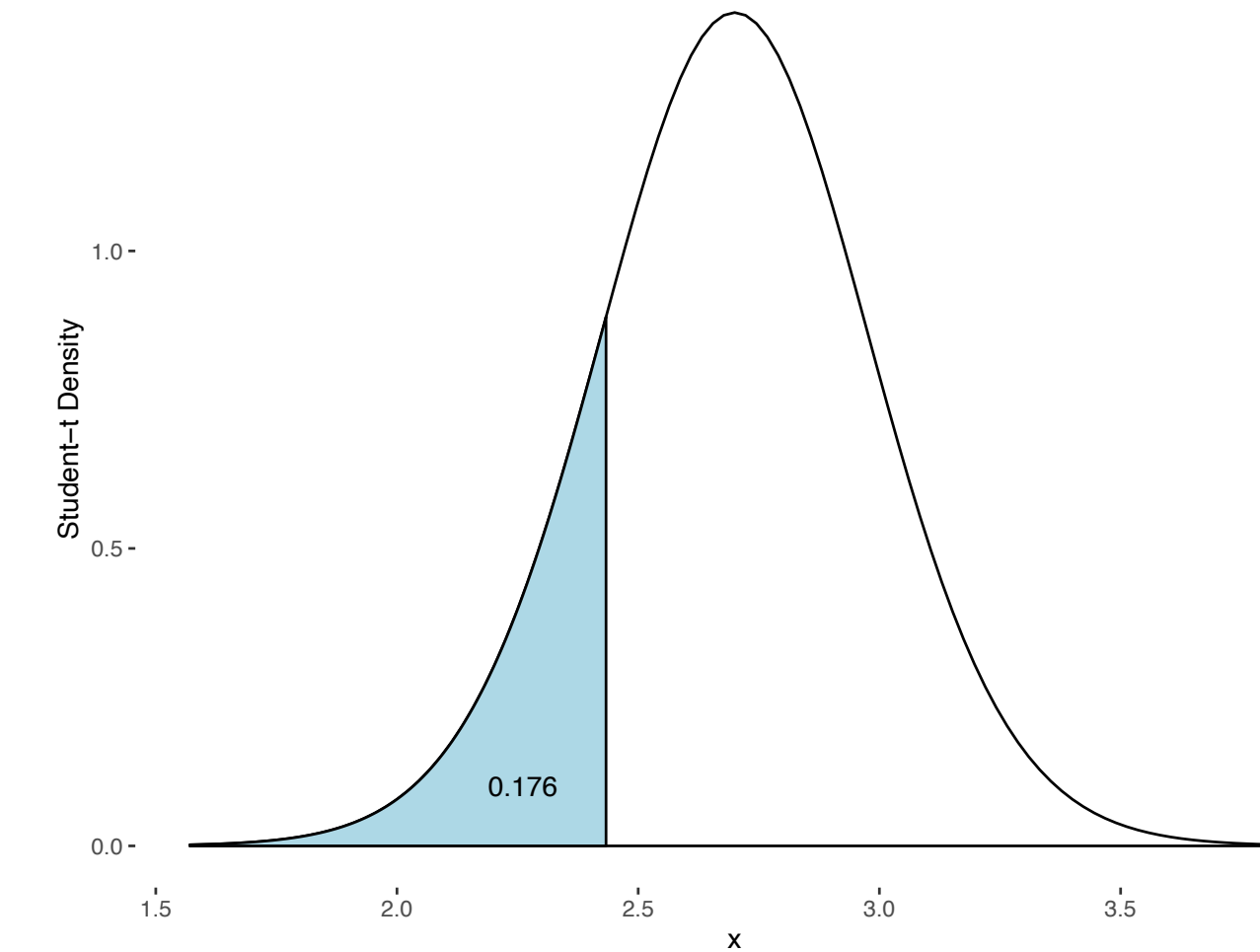
For ease of specifying student-t distribution parameters, here are functions for its four functions, just wrapping the base R versions:

```
dstudent_t <-  
  function(x, df, mu = 0, sigma = 1, log = FALSE) {  
    if (log) {  
      dt( (x - mu) / sigma, df = df, log = TRUE ) - log(sigma)  
    } else {  
      dt( (x - mu) / sigma, df = df ) / sigma  
    }  
  }  
  
pstudent_t <-  
  function(q, df, mu = 0, sigma = 1, lower.tail = TRUE, log.p = FALSE) {  
    pt( (q - mu) / sigma, df = df, lower.tail = lower.tail, log.p = log.p )  
  }  
  
qstudent_t <-  
  function(p, df, mu = 0, sigma = 1, lower.tail = TRUE, log.p = FALSE) {  
    qt( p, df = df, lower.tail = lower.tail, log.p = log.p ) * sigma + mu  
  }  
  
rstudent_t <-  
  function(n, df, mu = 0, sigma = 1) {  
    rt( n, df = df ) * sigma + mu  
  }
```

statistical tests, if **unknown**  $\sigma$ , can use sample standard deviation  $s$  and student's  $t$  distribution — *t-statistic*

$$H_0 : \mu = 2.7, H_A : \mu < 2.7$$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}, p = F_T(t)$$



```
s <- sd(x)

t <- ( x_bar - mu ) / ( s / sqrt(n) )
p <- pstudent_t(t, df = n - 1)
```

```
ggplot() +
  stat_function(fun = dstudent_t,
               args = list(df = n - 1, mu = mu, sigma = s / sqrt(n)),
               geom = "density",
               fill = "white",
               xlim = c(mu - 4 * s / sqrt(n),
                       mu + 4 * s / sqrt(n)) ) +
  stat_function(fun = dstudent_t,
               args = list(df = n - 1, mu = mu, sigma = s / sqrt(n)),
               geom = "density",
               fill = "lightblue",
               xlim = c(mu - 4 * s / sqrt(n), x_bar)) +
  annotate("segment", x = x_bar, xend = x_bar,
           y = 0, yend = dstudent_t(x_bar, n - 1, mu, s / sqrt(n))) +
  annotate("text", x = x_bar - 0.1, y = 0.1, hjust = 1,
           label = format(p, digits = 3)) +
  labs(y = "Student-t Density")
```

Given the null hypothesis, about  
17 percent of experiments have  $\bar{x} \leq 2.43$



statistical tests, (mis)interpreting test statistics, dichotomous tests, and a **warning**

Firstly,

$$P(D | H) \neq P(H | D)$$

$$P(H | D) = \frac{P(D | H)P(H)}{P(D | H)P(H) + P(D | \neg H)P(\neg H)}$$

Secondly,

**Dichotomous tests** are common in the literature. We typically read comparisons that ask whether the probability of obtaining a value more “extreme” than the sample mean from the null hypothesis, and compare this to an *arbitrary* p value of 0.05 (conventionally denoted  $\alpha$ ), rejecting the null hypothesis if this z or t value of smaller, not rejecting otherwise.

But — again — this threshold test is *arbitrary*. Consider that a difference between probabilities 0.049 and 0.051, for example, is not itself typically significant.

At least report the probability of obtaining the sample value given the selected null probability distribution, *not* just the result of some dichotomous test, and consider the probability of obtaining the sample value if the hypothesis is *not* true. That’s a future topic.

breakout session for group project kickoff

# References

**Blitzstein**, Joseph K., and Jessica Hwang. *Introduction to Probability*. Second edition. Boca Raton: Taylor & Francis, 2019.

**Gelman**, Andrew, Jennifer Hill, and Aki Ventari. *Regression and Other Stories*. S.l.: Cambridge University Press, 2020.

**Gelman**, Andrew. “The Problems With P-Values Are Not Just With P-Values.” *The American Statistician*, April 2016, 1–2.

**Harris**, Joseph. *Rewriting: How to Do Things with Texts*. Second edition. Logan: Utah State University Press, 2017.

**McShane**, Blakeley B., David Gal, Andrew Gelman, Christian Robert, and Jennifer L. Tackett. “Abandon Statistical Significance.” *The American Statistician* 73, no. sup1 (March 29, 2019): 235–45.