Agenda

1. Inference for a single proportion (practice)
2. Inference for a difference of two proportions

Warm-up: Inference for a Single Proportion

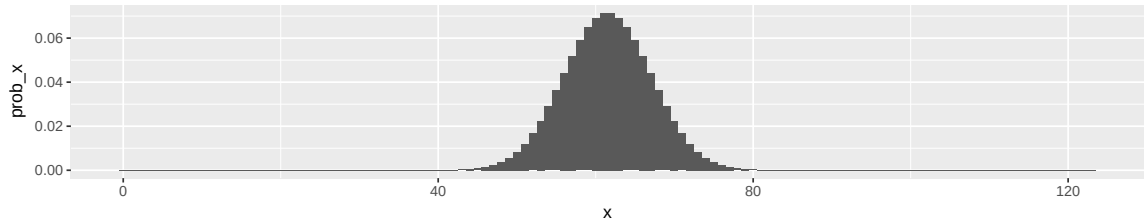
1. A company with a fleet of 1600 cars found that the emissions systems of 15 out of the 50 they tested failed to meet pollution control guidelines. Is this enough evidence that more than 20% of the fleet might be out of compliance? Test an appropriate hypothesis and state your conclusions.

2. A garden center wants to store leftover packets of vegetable seeds for sale the following spring, but the center is concerned that the seeds may not germinate at the same rate a year later. The manager finds a packet of last years green bean seeds and plants them as a test. Although the packet claims a germination rate of 92%, only 171 of 200 test seeds sprout. Is this evidence that the seeds have lost viability during a year in storage? Test an appropriate hypothesis and state your conclusion.

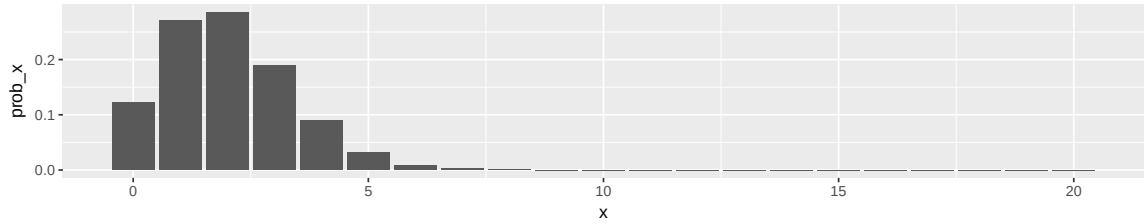
3. In 2009, a national vital statistics report indicated that about 3% of all births produced twins. Is the rate of twin births the same among very young mothers? Data from a large city hospital found that only 7 sets of twins were born to 469 teenage girls. Test an appropriate hypothesis and state your conclusion.

What Can Go Wrong? Most of the time, the null distribution for a proportion will be quite normal. In the coke example, the fit was excellent. But when the sample size is small of the probability of success is extreme, the binomial distribution will not look normal. Thus, before doing inference with the normal approximation, we will need to check what's called the *success/failure condition*.

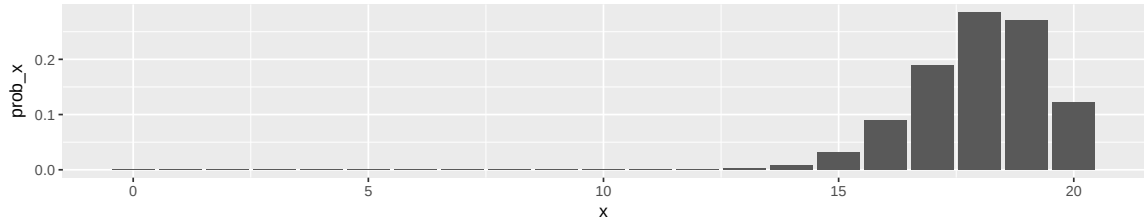
```
library(ggplot2)
n <- 123
p <- 0.50

x1  <- 0:n
df <- data.frame(x = x1, prob_x = dbinom(x1, n, p))

ggplot(df, aes(x, prob_x)) +
  geom_bar(stat = "identity")
```




```
n <- 20
p <- 0.10
```




```
n <- 20
p <- 0.90
```



Normal Approximation: Use statistical theory to *approximate* the null distribution.

1. Assumptions: independence, normality, $np > 10$ and $n(1-p) > 10$
2. Pros: uses familiar normal distribution, approximation is usually pretty good, possible to compute without computers (kind of)
3. Cons: requires more assumptions, not exact

Is the *success/failure conditon* satisfied for the problems in the warm-up?

Difference of two proportions Previously, we considered Ted Williams' batting average of .406 in 1941, which is unmatched in 75 years and counting. In 1994, Tony Gwynn of the San Diego Padres hit .394, but a strike by the player's union shortened the season after only 116 games. Thus, Gwynn accumulated 165 hits in 419 at-bats, whereas Williams had 185 hits in 456 at-bats. Let's assume that Gwynn had an unknown, fixed true batting average of p in 1994.

In many cases we will also want to make inferences about the difference between two proportions. Continuing the line of reasoning from above, let X be a binomial random variable that gives the number of hits that Williams will accrue in n_1 at-bats if his true batting average is p_1, and let Y be another binomial random variable that gives the number of hits that Gwynn will accrue in n_2 at-bats if his true batting average is p_2. Then we can define a new random variable

$$Z = \frac{X}{n_1} - \frac{Y}{n_2}$$

that gives the difference in their respective batting averages. Using linearity of expectation, we can compute the expected value of the difference:

$$E[Z] = E\left[\frac{X}{n_1} - \frac{Y}{n_2}\right] = \frac{1}{n_1} \cdot E[X] - \frac{1}{n_2} \cdot E[Y] = \frac{1}{n_1} \cdot n_1 p_1 - \frac{1}{n_2} \cdot n_2 p_2 = p_1 - p_2$$

and the variance:

$$\begin{aligned} Var[Z] &= Var\left[\frac{X}{n_1} - \frac{Y}{n_2}\right] = \frac{1}{n_1^2} \cdot Var[X] + \frac{1}{n_2^2} \cdot Var[Y] \\ &= \frac{1}{n_1^2} \cdot n_1 \cdot p_1(1-p_1) + \frac{1}{n_2^2} \cdot n_2 \cdot p_2(1-p_2) \\ &= \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} \end{aligned}$$

Just as before, this proves that the standard error is $SE_Z = SE_{\widehat{p_1 - p_2}} = \sqrt{SE_{\hat{p}_1}^2 + SE_{\hat{p}_2}^2}$. Once again, we'll typically use the normal approximation to the null distribution.

1. Using the normal approximation again, test the hypothesis that Williams and Gwynn had the same true batting averages in 1941 and 1994, respectively.

2. Since we are testing the hyopthesis that $p_1 = p_2$, it is more appropriate to use the *pooled* estimate of the standard error (see page 133). Perform this test.

3. Discuss the extent to which you think the performances of Williams and Gwynn were importantly different.

Practice: Difference of two proportions

1. The University of Michigan survey of consumers asked a sample of Americans Do you rate the governments economic policy as poor? Participants responded Yes or No. They also collected information on whether the participants only had high school diplomas (n = 140), or if they had college degrees (n = 87). Of those with high school diplomas, 70 said yes to the policy question. Of those with college degrees 30 said yes to the question. Test an appropriate hypothesis and state your conclusion.

2. A presidential candidate fears he has a problem with women voters. His campaign staff runs a poll to assess the situation. They randomly sample 300 men and 300 women, asking if they have a favorable impression of the candidate. They find that 59% of the men and 53% of the women have positive images of the candidate. Is there is difference in impression between all men and all women voters? Test an appropriate hypothesis and state your conclusion.

3. Researchers at the National Cancer Institute released the results of a study that investigated the effects of 827 dogs from homes where an herbicide was used on a regular basis, diagnosing malignant lymphoma in 473 of them. Of the 130 dogs from homes where no herbicides were used, only 19 were found to have lymphoma. Is there a different rate of cancer in dogs between homes that use the herbicide and homes that do not? Test an appropriate hypothesis and state your conclusion.