Chapter 8

Sampling Distributions

8.1 Introduction

To keep our discussion as general as possible, as the material in this subsection can be applied to many situations, we will let θ denote a population parameter. For example, θ could denote a population mean, a population variance, a population proportion, or some other model parameter, etc.

Recall: A point estimator $\hat{\theta}$ is a statistic that is used to estimate a population parameter θ. In this section, we're specifically interested in the following estimators:

$\bar{X} \to$ a point estimator for $E(X) = \mu_x$ (population mean)
$S^2 \to$ a point estimator for $var(X) = \sigma_x^2$ (population variance)
$S \to$ a point estimator for $sd(X) = \sigma_x$ (population standard deviation)
$\hat{p} \to$ a point estimator for $E(X) = p$ (population success probability for Bernoulli data)

Critical Point: A point estimator $\hat{\theta}$ is a statistic, so it depends on the sample of data $X_1, X_2, ..., X_n$.

- The data $X_1, X_2, ..., X_n$ come from the sampling process; e.g., different random samples will yield different data sets $X_1, X_2, ..., X_n$.
- In this light, because the sample values $X_1, X_2, ..., X_n$ will vary from sample to sample, the value of $\hat{\theta}$ will too. It therefore makes perfect sense to think about the distribution of $\hat{\theta}$ itself.

Terminology: The distribution of an estimator $\hat{\theta}$ is called its sampling distribution. A sampling distribution describes how the estimator $\hat{\theta}$ varies in repeated sampling.

Terminology: We say that $\hat{\theta}$ is an unbiased estimator of θ if and only if

$$E(\hat{\theta}) = \theta.$$

In other words, the expected value of the sampling distribution of $\hat{\theta}$ is equal to θ. Unbiasedness is a characteristic describing the center of a sampling distribution. This deals with accuracy.

Result: Mathematics shows that when $X_1, X_2, ..., X_n$ is a random sample,

$$E(\bar{X}) = E(X) = \mu_x$$
$$E(S^2) = var(X) = \sigma_x^2$$

That is, $\bar{X}$ and S^2 are unbiased estimators of their population analogues.

Goal: Not only do we desire to use point estimators $\hat{\theta}$ which are unbiased, but we would also like for them to have small variability. In other words, when $\hat{\theta}$ "misses" θ, we would like for it to "not miss by much." This deals with precision.

Terminology: The standard error of a point estimator $\hat{\theta}$ is equal to

$$se(\hat{\theta}) = \sqrt{var(\hat{\theta})}.$$

In other words, the standard error is equal to the standard deviation of the sampling distribution of $\hat{\theta}$. An estimator's standard error measures the amount of variability in the point estimator $\hat{\theta}$. Therefore,

$$\text{smaller } \ se(\hat{\theta}) \Leftrightarrow \hat{\theta} \ \text{ more precise.}$$

Main point: As discussed, accuracy and precision are the two main mathematical characteristics that arise when evaluating the quality of a point estimator $\hat{\theta}$. We desire point estimators $\hat{\theta}$ which are unbiased (perfectly accurate) and have small variance (highly precise).

Illustration: In Example 7.17 we compared the MOM and MLE estimates to data from the Uniform($a = 0$,$b = 3$) distribution where we treated b as unknown.

- The estimators for unknown b were $\hat{b} = 2\bar{X}$ (MOM) and $\hat{b} = X_{(n)}$ (MLE).
- The empirical expected value of these estimators, calculated via simulation, were $E(\hat{b}) = 3.001$ (MOM) and $E(\hat{b}) = 2.997$ (MLE).
- The empirical standard error of these estimators, calculated via simulation, were $se(\hat{b}) = \sqrt{0.003}$ (MOM) and $se(\hat{b}) = \sqrt{0.0000077}$ (MLE).
- The empirical distribution of these estimators, estimated via simulation, were bell-shaped (MOM) and heavily left skewed (MLE).

We are interested in the expected values, standard errors and even the distribution of estimators. We were able to explore them via simulation in the previous chapter, however this approach isn't helpful with real data. In this chapter, we discuss several approaches for discussing these topics with real data provided by theoretical mathematics.

8.2 Notation

Below is a table of how we denote many parameters and statistics in this chapter. Seeing these shorthand notations as what they represent will be more meaningful than seeing them as a random collection of letters and Greek letters. We know many of them already, and we will define new entries in the following sections.

Letter	Statistic or Parameter?	What does it represent?
n	.....	Sample size
$\overline{x}$	Statistic	Sample mean
μ_x	Parameter	The average of the population
$\mu_{\widehat{p}}$	Parameter	Mean of all possible sample proportions
$\mu_{\overline{x}}$	Parameter	Mean of all possible sample means
p	Parameter	Population proportion
$\widehat{p}$	Statistic	Sample proportion
s	Statistic	Sample standard deviation
s^2	Statistic	Sample variance
σ_x	Parameter	Population standard deviation
σ_x^2	Parameter	Population variance
$\sigma_{\widehat{p}}$	Parameter	Standard deviation of all possible sample proportions
$\sigma_{\overline{x}}$	Parameter	Standard deviation of all possible sample means

8.3 The Central Limit Theorem (CLT)

8.3.1 The Sampling Distribution of the Sample Mean

Definition 8.1. (CLT 1: Gaussian Data) Suppose that $X_1, X_2, ..., X_n$ is a random sample from a Gaussian(μ,σ^2) distribution. The sample mean $\bar{X}$ has the following **sampling distribution**:

$$\bar{X} \sim \text{Gaussian}\left(\underbrace{\mu_x}_{\mu_{\bar{x}}}, \underbrace{\frac{\sigma_x^2}{n}}_{\sigma_{\bar{x}}}\right).$$

This result reminds us that

$$E(\bar{X}) = \mu_{\bar{x}} = \mu_x.$$

That is, the sample mean $\bar{X}$ is an unbiased estimator of the population mean μ_x.

This result also shows that the standard error of $\bar{X}$ (as a point estimator) is

$$se(\bar{X}) = \sigma_{barx} = \sqrt{var(\bar{X})} = \sqrt{\frac{\sigma_x^2}{n}}.$$

We explore this by graphing a Gaussian($\mu = 0, \sigma = 1$) population distribution and the sampling distributions for $\bar{X}$ for $n = 2, 10, 15, 30, 50$ in Figure 8.3.1, created with the R code.

```
> x<-seq(-4,4,0.01)
> ggdat<-data.frame(x=x,
+                   f.pop=dnorm(x,0,1),
+                   f.samp2=dnorm(x,0,1/sqrt(2)),
+                   f.samp10=dnorm(x,0,1/sqrt(10)),
+                   f.samp15=dnorm(x,0,1/sqrt(15)),
+                   f.samp30=dnorm(x,0,1/sqrt(30)),
+                   f.samp50=dnorm(x,0,1/sqrt(50)))
> ggplot(data=ggdat,aes(x=x,y=f.pop))+
+   geom_line(color="black")+
+   geom_line(aes(y=f.samp2,color="n=2"))+
+   geom_line(aes(y=f.samp10,color="n=10"))+
+   geom_line(aes(y=f.samp15,color="n=15"))+
+   geom_line(aes(y=f.samp30,color="n=30"))+
+   geom_line(aes(y=f.samp50,color="n=50"))+
+   geom_hline(yintercept = 0)+
+   theme_bw()+
+   xlab(bquote(bar(x)))+
+   ylab("Density")+
+   ggtitle("Sampling Distributions",
+         subtitle="Population distribution superimposed in black")+
+   scale_color_discrete("",breaks=c("n=2","n=10","n=15","n=30","n=50"))
```

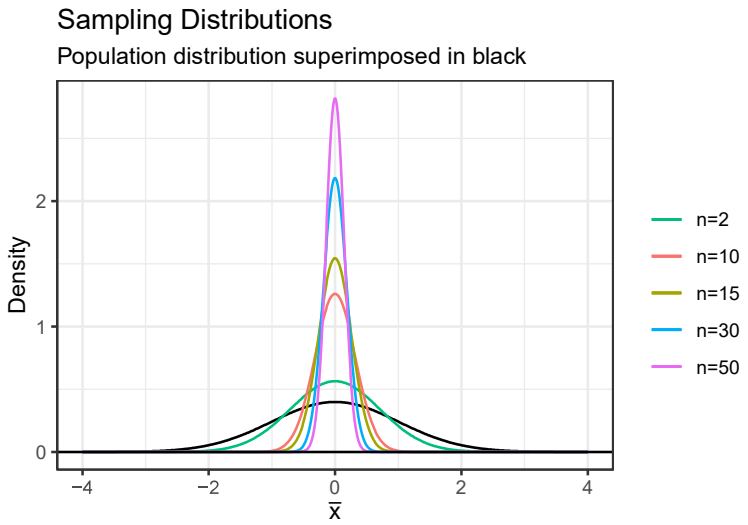

Figure 8.3.1: A Gaussian($\mu = 0, \sigma = 1$) population distribution (in black) and the sampling distributions for $\bar{X}$ for $n = 2, 10, 15, 30, 50$, which are Gaussian($\mu_{\bar{x}} = 0, \sigma_{\bar{x}} = 1/\sqrt{n}$).

Curiosity: Why are the sampling distributions narrower than the population distribution? The sampling distributions are narrower because they have standard error smaller than the population standard deviation because

$$\sigma_x \geq \frac{\sigma_x}{\sqrt{n}},$$

for $\sigma > 0$ and $n \geq 1$.

We can see this in our law of large numbers plots. For example, in Figure 8.3.2, we see two examples of simulated data from the Gaussian($\mu_x = 0, sigma_x = 1$) showing that as the sample size grows, the sample mean approaches the population mean and the variation around the true population mean decreases..

```
> x<-rnorm(5000,0,1)
> xbar<-cumsum(x)/1:5000
> ggdat<-data.frame(xbar=xbar)
> g1<-ggplot(data=ggdat, aes(x=1:5000, y=xbar))+
+   geom_line()+
+   geom_hline(yintercept = 0,color="red",linetype="dashed")+
+   theme_bw()+
+   ylab(bquote(bar(x)))+
+   xlab("Observation")+
+   ggtitle("Law of Large Numbers",
+         subtitle=bquote(bar(X)~"approaches E(X) ="~mu==0~"as n increases"))
> x<-rnorm(5000,0,1)
> xbar<-cumsum(x)/1:5000
> ggdat<-data.frame(xbar=xbar)
> g2<-ggplot(data=ggdat, aes(x=1:5000, y=xbar))+
+   geom_line()+
+   geom_hline(yintercept=0,color="red",linetype="dashed")+
+   theme_bw()+
+   ylab(bquote(bar(x)))+
+   xlab("Observation")+
+   ggtitle("Law of Large Numbers",
+         subtitle=bquote(bar(X)~"approaches E(X) ="~mu==0~"as n increases"))
> grid.arrange(g1,g2,ncol=2)
```

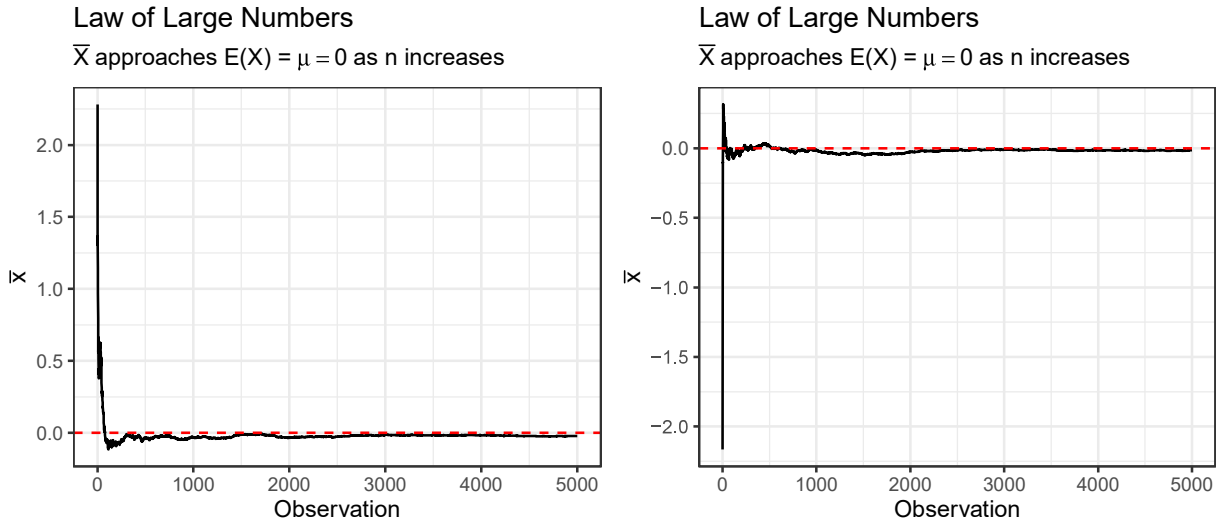

Figure 8.3.2: Two examples of simulated data from the Gaussian($\mu_x = 0, sigma_x = 1$) showing that as the sample size grows, the sample mean approaches the population mean and the variation around the true population mean decreases.

In Figure 8.3.3, we superimpose 50 such cases of simulated data. We see that the band of cumulative estimators centers over the population mean and gets narrower over time. This indicates that the estimator is accurate and less varied, thus more precise, as the sample size gets larger.

```
> x<-rnorm(5000,0,1)
> xbar<-cumsum(x)/1:5000
> ggdat<-data.frame(xbar=xbar)
> g<-ggplot(data=ggdat, aes(x=1:5000, y=xbar))+
+   geom_line()+
+   theme_bw()+
+   ylab(bquote(bar(x)))+
+   xlab("Observation")+
+   ggtitle("Law of Large Numbers",
+         subtitle=bquote(bar(X)~"approaches E(X) ="~mu==0~"as n increases"))
> for(i in 2:50){
+   x<-rnorm(5000,0,1)
+   xbar<-cumsum(x)/1:5000
+   ggdat<-data.frame(xbar=xbar)
+   g<-g+geom_line(data=ggdat,aes(x=1:5000,y=xbar),color=i)
+ }
> g
```

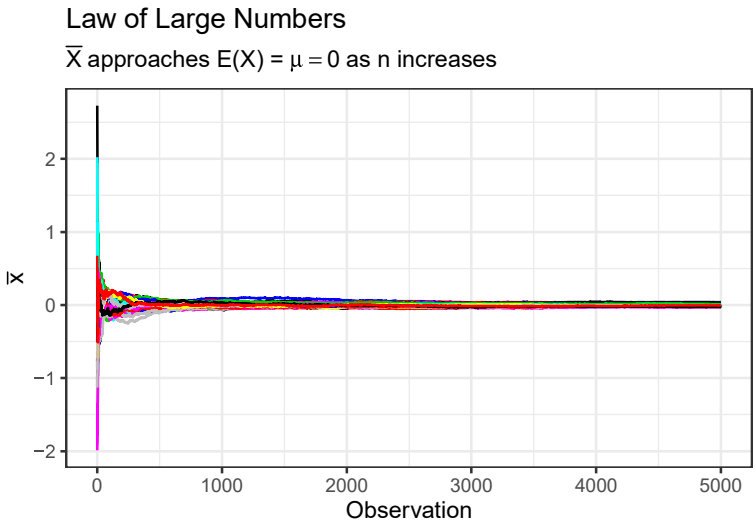

Figure 8.3.3: Fifty examples of simulated data from the Gaussian($\mu_x = 0, sigma_x = 1$) showing that as the sample size grows, the sample mean approaches the population mean and the variation around the true population mean decreases.

Definition 8.2. (CLT 2: Model Agnostic) Suppose that $X_1, X_2, \ldots X_n$ is a random sample with mean μ_x and known standard deviation $\sigma_x < \infty$. As our sample size increases,

$$\bar{X} \sim \mathcal{AG}(\mu_{\bar{x}} = \mu_x, \sigma_{\bar{x}} = \sigma_x/\sqrt{n}).$$

where $\mathcal{AG}$ denotes "approximately Gaussian." This implies that

$$Z = \frac{\bar{X} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{X} - \mu_x}{\sigma_x/\sqrt{n}} \sim \mathcal{AG}(0, 1),$$

We say that the distribution of Z is approximately the standard Gaussian distribution, the Gaussian distribution with mean zero and standard error one; and the distribution of $\bar{X}$ is approximately Gaussian with mean $\mu_{\bar{x}}$ and standard error $\sigma_{\bar{x}}$.

Remark: Note that this result is very powerful! The Central Limit Theorem (CLT) states that sample averages will be approximately Gaussian distributed even if the underlying population distribution, $f_X(x)$, is not!

Remark: Note that there is a distinction to be made between the theoretical sampling distribution and our approximation. The sampling distribution may not be Gaussian (it may be Binomial, or Gamma, or some other distribution), but as the sample size increases the Gaussian distribution will approximate the theoretical sampling distribution well regardless of the underlying theoretical distribution.

Q: How good is the approximation?
A: Since the CLT only offers an approximate sampling distribution for $\bar{X}$, one might naturally wonder exactly how good the approximation is. In general, the goodness of the approximation jointly depends on

- Sample size – the larger the sample size, the better the approximation is; e.g., the more Gaussian the sampling distribution is.
- Sample size – as the sample size increases the variation in sample mean observations decreases; e.g., the more narrow the sampling distribution is.
  - Recall our discussions about the law of large numbers; because the sample mean approaches the population mean and varies less as our sample size increases, our measure of variability for $\bar{X}$ should decrease as n increases as well. This is explicit in the formula for $\sigma_{\bar{x}}$ as we divide by $\sqrt{n}$.
- Symmetry in the underlying population distribution $f_X(x)$ – the more symmetric $f_X(x)$ is, the better the approximation.

When talking about the sampling distribution of the sample mean we say that the sampling distribution is approximately Gaussian if either of the following are true

- $n \geq 30$
- the population distribution is Gaussian (regardless of n)

It is important to note that $n \geq 30$ is a guideline that works well for many distributions, even highly skewed distributions. Some population distributions, particularly those that are more symmetric, don't require n to be as large. You will explore this more thoroughly for a handful of distributions as an exercise.

Example 8.3. Facebook's first quarter report, described in a May 6, 2016 New York Times article (Stewart, 2016), reported that the average user of Facebook, Instagram and Messenger platforms (not including WhatsApp) used these technologies for 50 minutes per day, up from around 40 minutes just two years prior.

Suppose that users of Facebook, Instagram and Messenger platforms (not including WhatsApp) used these technologies for 50 minutes a day, on average, with a standard deviation of 25 minutes.

Q: What is the probability that a randomly selected user spends less than 45 minutes per day on the Facebook, Instagram and Messenger platforms?

A: This is a probability about a randomly selected person, so we would use the population distribution to answer this question. Unfortunately, this was not provided to us so we do not have enough information to answer the question.

Q: What is the sampling distribution of the sample mean from a sample of $n = 100$?

A: We can find the sampling distribution of the sample mean using the formulas

$$\mu_{\bar{x}} = \mu_x = 50 \text{ minutes}$$
$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{25}{\sqrt{100}} = 2.5 \text{ minutes.}$$

Despite not knowing the population distribution, we know that $\bar{X}$ will be approximately Gaussian distributed because the sample size, $n = 100$, is greater than thirty which satisfies the benchmark.

Q: What is the probability that a random sample of 100 users spend more than 45 minutes per day on the Facebook, Instagram and Messenger platforms, on average?

A: This is a question about a sample so we use the sampling distribution to answer this question.

$$P(\bar{X} \leq 45) = 0.02275013$$

As calculated in R.

```
> m.xbar<-50
> se.xbar<-25/sqrt(100)
> pnorm(q=45,mean=m.xbar,sd=se.xbar)
[1] 0.02275013
```

Remark: It is important to note that we use the mean and variance of the sampling distribution when asking for this probability.

Which is the area under the sampling distribution as depicted in Figure 8.3.4, which is created with the following R code.

```
> ggdat<-data.frame(x=seq(40,60,0.01),
+                   f1=dnorm(x=seq(40,60,0.01),mean=m.xbar,sd=se.xbar))
> g1<-ggplot(data=ggdat,aes(x=x))+
+   geom_line(aes(y=f1))+
+   geom_ribbon(data=subset(ggdat,x<=45),aes(ymax=f1),ymin=0,
+               fill="red",colour=NA,alpha=0.5)+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab(bquote("Minutes per day"~(bar(X))))+
+   ylab(bquote(f[bar(x)](bar(x))))+
+   ggtitle("Gaussian PDF",
+        subtitle=bquote(P(bar(X)<=45)~"for"~mu[bar(X)]==50*","~sigma[bar(X)]==2.5))
```

```
> ggdat<-data.frame(x=seq(40,60,0.01),
+                   F1=pnorm(q=seq(40,60,0.01),mean=m.xbar,sd=se.xbar))
> ggdat.highlight<-data.frame(x=45,
+                             y=pnorm(q=45,mean=m.xbar,sd=se.xbar))
> g1.cdf<-ggplot(data=ggdat,aes(x=x))+
+   geom_line(aes(y=F1))+
+   geom_point(data=ggdat.highlight,aes(x=x,y=y),color="red")+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab(bquote("Minutes per day"~(bar(X))))+
+   ylab(bquote(F[bar(x)](bar(x))))+
+   ggtitle("Gaussian CDF",
+        subtitle=bquote(P(bar(X)<=45)~"for"~mu[bar(X)]==50*","~sigma[bar(X)]==2.5))
> grid.arrange(g1,g1.cdf,ncol=2)
```

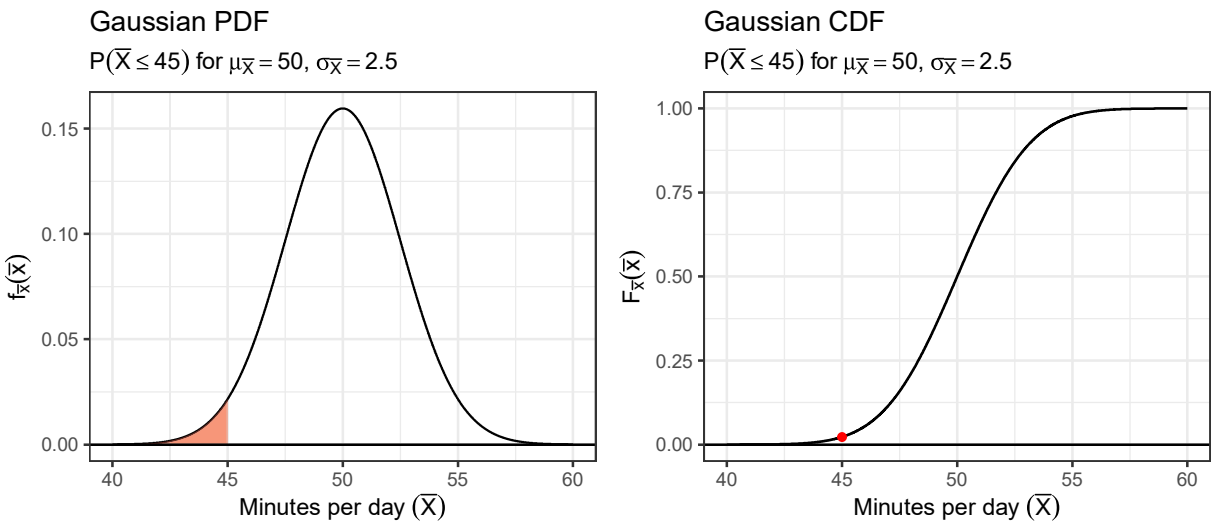

Figure 8.3.4: The sampling distribution, Gaussian PDF (left) for $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$ with area under the PDF shaded to the right of 45 representing $P(\bar{X} > 45)$ and CDF (right) with $P(\bar{X} \leq 45)$ highlighted with a red point.

Example 8.4. A chemist is studying the degradation behavior of vitamin B6 in a multivitamin. The chemist selects a random sample of $n = 36$ multivitamin tablets, and for each tablet, counts the number of days until the B6 content falls below the FDA requirement. The manufacturer suggests that X_1, X_2, ..., $X_{36} \sim$ Poisson($\lambda = 50$) for the thirty six tablets.

Q: What is the probability that the number of days until the B6 content falls below the FDA requirement for vitamin one, X_1, will exceed 52 days? That is, what is $P(X_1 > 52)$?

A: This is a probability about a randomly selected tablet so we use the population distribution to answer this question.

$$\begin{aligned} P(X_1 > 52) &= 1 - P(X_1 \leq 52) \\ &= 0.3542 \end{aligned}$$

This probability can be calculated in R as follows.

```
> 1-ppois(q=52,lambda=50)
[1] 0.354166
> ppois(q=52,lambda=50,lower.tail=FALSE)
[1] 0.354166
```

This probability is visualized in Figure 8.3.5, created with the following R code.

```
> ggdat<-data.frame(x=(20:80),
+                   f1=dpois(x=(20:80),lambda=50),
+                   F1=ppois(q=(20:80),lambda=50))
> ggdat.highlight<-data.frame(x=52:80,
+                             f1=dpois(x=52:80,lambda=50))
> g1<-ggplot(data=ggdat,aes(x=x))+
+   geom_linerange(aes(ymax=f1), ymin=0)+
+   geom_linerange(data=ggdat.highlight,aes(ymax=f1),ymin=0,color="red")+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("Days to Degradation (X)")+
+   ylab(bquote(f[x](x)))+
+   ggtitle("Poisson PMF",subtitle=bquote(P(X>52)~"for"~lambda==50))
> ggdat.openpoints<-data.frame(x=ggdat$x,
+                              y=ppois(ggdat$x-1,lambda=50))
> ggdat.closedpoints<-data.frame(x=ggdat$x,
+                                y=ppois(ggdat$x,lambda=50))
> ggdat.highlight<-data.frame(x=52,
+                             y=ppois(52,lambda=50))
> g1.CDF<-ggplot(data=ggdat, aes(x = x, y = F1)) +
+   geom_step()+
+   geom_point(data = ggdat.openpoints, aes(x = x, y = y),shape=1) +
+   geom_point(data = ggdat.closedpoints, aes(x = x, y = y)) +
+   geom_point(data = ggdat.highlight, aes(x = x, y = y),color="red") +
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("Days to Degradation (X)")+
+   ylab(bquote(F[x](x)))+
+   ggtitle("Poisson CDF",subtitle=bquote(P(X<=52)~"for"~lambda==50))
> grid.arrange(g1,g1.CDF,ncol=2)
```

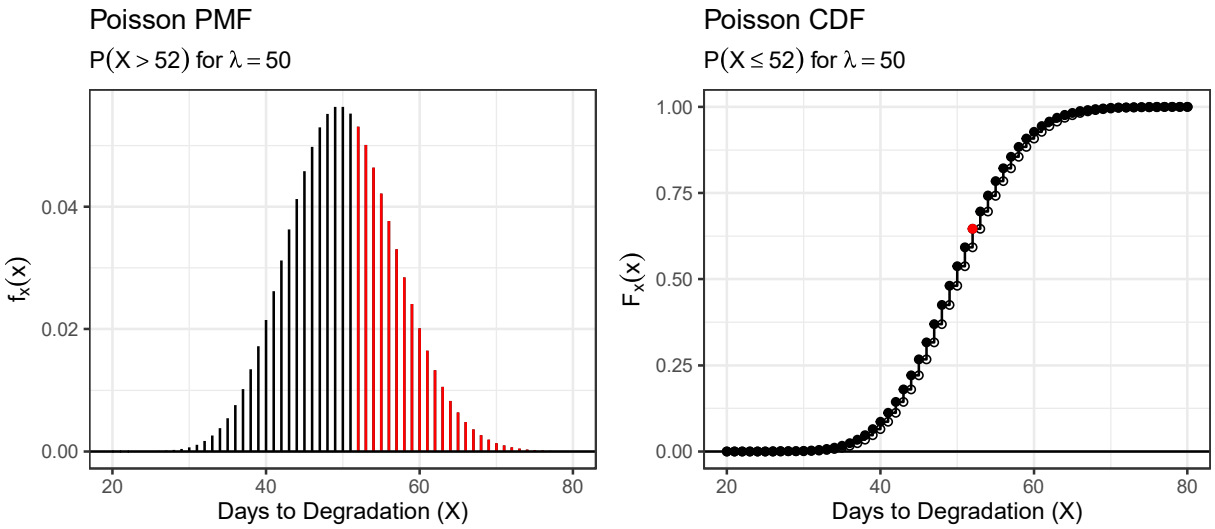

Figure 8.3.5: Poisson PMF (left) for $\lambda = 50$ with the bars from $X = 53$ to the right highlighted in red which represents $P(X > 52)$ and Poisson CDF (right) with $P(X \leq 2)$ highlighted.

Q: Suppose a consumer group wanted to check the statement that the multivitamins last, on average, 50 days before they fall below the FDA requirement. The consumer group bought a bottle, checked each pill daily and found the average number of days before the vitamins fell below the FDA requirement of 52 days. What is the approximate probability that the average number of days it takes the 36 multivitamins to fall below the FDA requirement, $\bar{X}$ will exceed 52 days? That is, what is $P(\bar{X} > 52)$?

A: This is a question about a sample so we use the approximate sampling distribution to answer this question.

Since we know the measurements follow the Poisson distribution with mean, $E(X) = \lambda = 50$, and variance, $var(X) = \lambda = 50$, the sampling distribution is calculated as follows

$$E(X) = \lambda = 50 \longrightarrow E(\bar{X}) = \mu_{\overline{X}} = \mu_X = 50$$

$$var(X) = \lambda = 50 \longrightarrow var(\bar{X}) = \sigma_{\overline{X}} = \frac{\sigma_x}{n} = \sqrt{\frac{50}{36}} = 1.1785.$$

We know that $\bar{X}$ will be approximately Gaussian distributed because the sample size, $n = 36$, is greater than thirty which satisfies our benchmark. Thus,

$$\bar{X} \sim \mathcal{AG}(\mu_{\bar{x}} = 50, \sigma_{\bar{x}} = 1.1785).$$

We can evaluate the probability of interest as follows.

$$\begin{aligned} P(\bar{X} > 52) &= 1 - P(\bar{X} \leq 52) \\ &= 0.0448 \end{aligned}$$

```
> m.xbar<-50
> se.xbar<-sqrt(50/36)
> 1-pnorm(q=52,mean=m.xbar,sd=se.xbar)
```

```
[1] 0.04484301
> pnorm(q=52,mean=m.xbar,sd=se.xbar,lower.tail=FALSE)
[1] 0.04484301
```

Which is the area under the sampling distribution as depicted in Figure 8.3.6 which is created using the following R.

```
> ggdat<-data.frame(x=seq(45,55,0.01),
+                   f1=dnorm(x=seq(45,55,0.01),mean=m.xbar,sd=se.xbar))
> g1<-ggplot(data=ggdat,aes(x=x))+
+   geom_line(aes(y=f1))+
+   geom_ribbon(data=subset(ggdat,x>=52),aes(ymax=f1),ymin=0,
+               fill="red",colour=NA,alpha=0.5)+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab(bquote("Days to Degradation"~(bar(X))))+
+   ylab(bquote(f[bar(x)](bar(x))))+
+   ggtitle("Gaussian PDF",
+        subtitle=bquote(P(X>=52)~"for"~mu==50*","~sigma==1.179))
> ggdat<-data.frame(x=seq(45,55,0.01),
+                   F1=pnorm(q=seq(45,55,0.01),mean=m.xbar,sd=se.xbar))
> ggdat.highlight<-data.frame(x=52,
+                             y=pnorm(q=52,mean=m.xbar,sd=se.xbar))
> g1.cdf<-ggplot(data=ggdat,aes(x=x))+
+   geom_line(aes(y=F1))+
+   geom_point(data=ggdat.highlight,aes(x=x,y=y),color="red")+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab(bquote("Days to Degradation"~(bar(X))))+
+   ylab(bquote(F[bar(x)](bar(x))))+
+   ggtitle("Gaussian CDF",
+        subtitle=bquote(P(X<=52)~"for"~mu==50*","~sigma==1.179))
> grid.arrange(g1,g1.cdf,ncol=2)
```

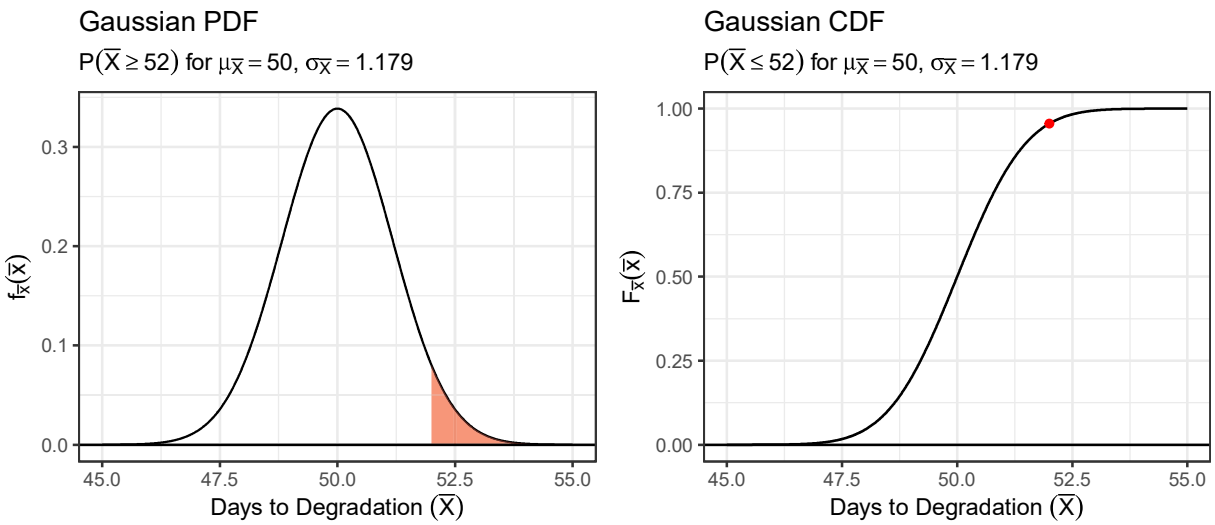

Figure 8.3.6: The sampling distribution, Gaussian PDF for $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$ (left) with area under the PDF shaded to the right of 52 representing $P(\bar{X} > 52)$ calculated in Example 8.4. The Gaussian CDF for $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$ (right) with a point at $x = 52$ representing $P(\bar{X} \leq 52)$

Curiosity: Does $\bar{X} = 52$ seem like a value you would expect to see from this distribution? Recall that this probability was computed under the assumption that $\mu_{\bar{x}} = \sigma^2_{\bar{X}} = 50$, as specified by the manufacturer. Therefore, if this claim is true, we would expect to see a value of $\bar{X}$ around the center of its sampling distribution. This case is a close call – some might see this as close enough to the center, while some may see this as somewhat unusual (out in the tail) which would be more consistent with a value of $\mu_{\bar{x}}$ that is larger than 50 days.

Q: Suppose the manufacturer of the multivitamins wants to provide their marketing team with a statement about the degradation time. To investigate the appropriate statement to make about the degradation time for bottles of 36 tablets they ask which observation, x, they would need to observe so that $P(\bar{X} < x) \approx 0.01$?

A: This is asking for the 1st percentile of the distribution of $\bar{X}$, the observation that is larger than only one percent of observable values. This requires us to use the inverse CDF as follows.

$$F^{-1}(0.01) = 47.26$$

This is calculated in `R` as follows.

```
> m.xbar<-50
> se.xbar<-sqrt(50/36)
> qnorm(p=0.01,mean=m.xbar,sd=se.xbar)
[1] 47.25837
```

We want to find the value x so that the area under the sampling distribution to the left of x is 0.01 as depicted in Figure 8.3.7, created with the following `R` code.

```
> ggdat<-data.frame(x=seq(45,55,0.01),
+                   f1=dnorm(x=seq(45,55,0.01),
```

```
+                     mean=m.xbar,sd=se.xbar))
> g1<-ggplot(data=ggdat,aes(x=x))+
+   geom_line(aes(y=f1))+
+   geom_ribbon(data=subset(ggdat,x<=47.25837),aes(ymax=f1),ymin=0,
+               fill="red",colour=NA,alpha=0.5)+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab(bquote("Days to Degradation"~(bar(X))))+
+   ylab(bquote(f[bar(x)](bar(x))))+
+   ggtitle("Gaussian PDF",
+         subtitle=bquote(F^(-1)*(0.01)~"for"~mu[bar(X)]==50*","~sigma[bar(X)]==1.179))
> ggdat<-data.frame(x=seq(45,55,0.01),
+                   F1=pnorm(q=seq(45,55,0.01),mean=m.xbar,sd=se.xbar))
> ggdat.highlight<-data.frame(x=qnorm(p=0.01,mean=m.xbar,sd=se.xbar),
+                             y=pnorm(q=qnorm(p=0.01,
+                                             mean=m.xbar,sd=se.xbar),
+                                     mean=m.xbar,sd=se.xbar))
> g1.cdf<-ggplot(data=ggdat,aes(x=x))+
+   geom_line(aes(y=F1))+
+   geom_point(data=ggdat.highlight,aes(x=x,y=y),color="red")+
+   geom_hline(yintercept=0)+
+   geom_hline(yintercept=0.01,linetype="dashed")+
+   theme_bw()+
+   xlab(bquote("Days to Degradation"~(bar(X))))+
+   ylab(bquote(F[bar(x)](bar(x))))+
+   ggtitle("Gaussian CDF"
+         subtitle=bquote(F^(-1)*(0.01)~"for"~mu[bar(X)]==50*","~sigma[bar(X)]==1.179))
> grid.arrange(g1,g1.cdf,ncol=2)
```

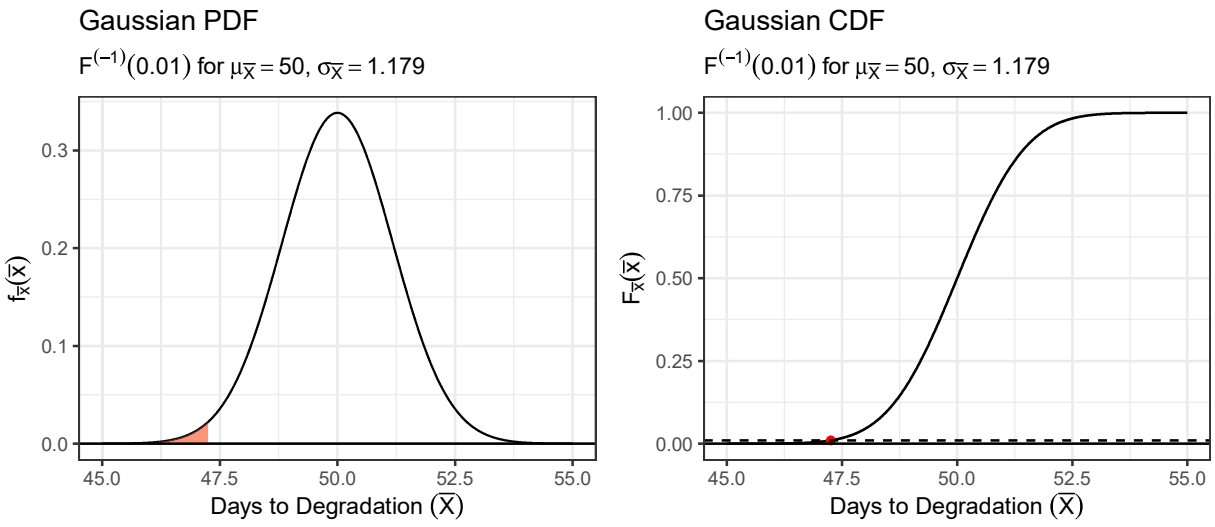

Figure 8.3.7: The sampling distribution, Gaussian PDF for $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$ with area representing $P(\bar{X} < x) \approx 0.01$ calculated in Example 8.4. The value of interest is x, the value on the x-axis where the area ends.

This can be calculated and visualized in R as follows.

Q: Suppose the manufacturer wants to bottle the multivitamins so that the probability that the average degradation time is less than 49.5 is approximately 0.01. That is how many tablets does the manufacturer need to observe so that $P(\bar{X} < 49.5) \approx 0.01$?

A: This question is more challenging, we cannot ask R for this value directly. What happens to the sampling distribution as n changes can be seen in Figure 8.3.8 created with the following R code.

```
> ggdat<-data.frame(x=seq(45,55,0.01),
+                   f30=dnorm(seq(45,55,0.01),mean=m.xbar,sd=sqrt(50/30)),
+                   f36=dnorm(seq(45,55,0.01),mean=m.xbar,sd=sqrt(50/36)),
+                   f50=dnorm(seq(45,55,0.01),mean=m.xbar,sd=sqrt(50/50)),
+                   f100=dnorm(seq(45,55,0.01),mean=m.xbar,sd=sqrt(50/100)),
+                   f500=dnorm(seq(45,55,0.01),mean=m.xbar,sd=sqrt(50/500)),
+                   f1000=dnorm(seq(45,55,0.01),mean=m.xbar,sd=sqrt(50/1000)))
> ggplot(data=ggdat,aes(x=x))+
+   geom_line(aes(y=f30,color="n=30"))+
+   geom_line(aes(y=f36,color="n=36"))+
+   geom_line(aes(y=f50,color="n=50"))+
+   geom_line(aes(y=f100,color="n=100"))+
+   geom_line(aes(y=f500,color="n=500"))+
+   geom_line(aes(y=f1000,color="n=1000"))+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab(bquote("Days to Degradation"~(bar(X))))+
+   ylab(bquote(f[bar(x)](bar(x))))+
+   ggtitle("Gaussian PDF",
+           subtitle=bquote("Varying"~n~"for"~mu[bar(X)]==50*","~sigma[bar(X)]==sqrt(50/n)))+
+   scale_color_discrete("")
```

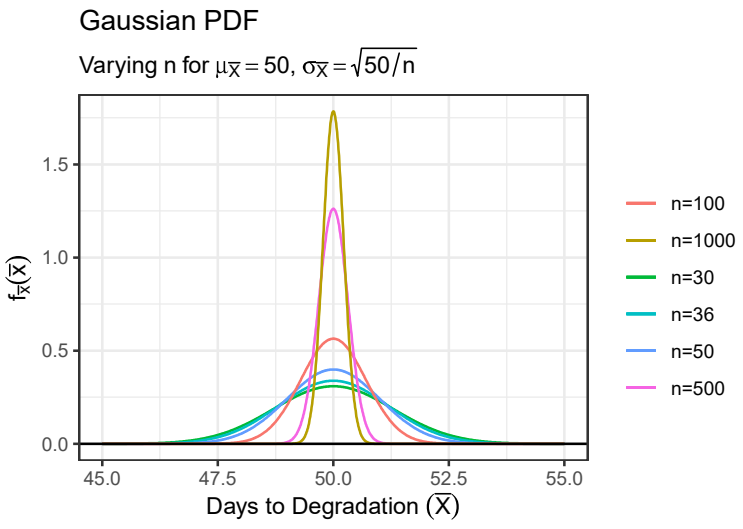

Figure 8.3.8: The sampling distribution, Gaussian PDF for $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$ with $n = 30, 36, 100, 500, 1000$ from Example 8.4.

We want to find n so that the area to the left of 49.5 under the sampling distribution for $\bar{X}$ is 0.01. To answer this question we must consider the standard Gaussian transformation. We want to find n such that

$$P(\bar{X} < 49.5) \approx P\left(Z_n < \frac{49.5 - 50}{\sqrt{50/n}}\right) \approx 0.01.$$

To do this we need to find the 1st percentile of the standard Gaussian distribution,

```
> qnorm(p=0.01,mean=0,sd=1)
[1] -2.326348
```

and solve

$$\begin{aligned}
Z_n = \frac{49.5 - 50}{\sqrt{50/n}} &= -2.3263 \\
49.5 - 50 &= -2.3263\sqrt{50/n} && \textbf{[Multiply both sides by } \sqrt{50/n}\textbf{]} \\
\frac{49.5 - 50}{-2.3263} &= \sqrt{50/n} && \textbf{[Divide both sides by } -2.3263\textbf{]} \\
\left(\frac{49.5 - 50}{-2.3263}\right)^2 &= 50/n && \textbf{[Square both sides]} \\
n\left(\frac{49.5 - 50}{-2.3263}\right)^2 &= 50 && \textbf{[Multiply both sides by } n\textbf{]} \\
n &= \frac{50}{\left(\frac{49.5-50}{-2.3263}\right)^2} && \textbf{[Divide both sides by } \left(\tfrac{49.5-50}{-2.3263}\right)^2\textbf{]} \\
&= 1082.334
\end{aligned}$$

It follows that $n \approx 1082$.

8.3.2 The Sampling Distribution of the Sample Mean – unknown σ

Suppose that $X_1, X_2, ..., X_n$ is a random sample from a Gaussian(μ_x, σ_x^2) distribution. Recall that Theorem 8.1 says the sample mean $\bar{X}$ has the following sampling distribution:

$$\bar{X} \sim \text{Gaussian}\left(\underbrace{\mu_x}_{\mu_{\bar{x}}}, \underbrace{\frac{\sigma_x^2}{n}}_{\sigma_{\bar{x}}}\right).$$

If we standardize $\bar{X}$, we obtain

$$Z = \frac{\bar{X} - \mu_x}{\sigma_x/\sqrt{n}} \sim \text{Gaussian}(\mu_z = 0, \sigma_z = 1).$$

If we replace the population standard deviation σ_x with the sample standard deviation S, we get a new sampling distribution:

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t(n-1).$$

a t distribution with degrees of freedom $v = n - 1$.

Recall: The t pdf has the following characteristics:

- It is continuous and symmetric about 0 (just like the standard Gaussian PDF).
- It is indexed by a value v called the degrees of freedom. In practice, v is often an integer (related to the sample size).
- As $v \to \infty$, $t(v) \to \text{Gaussian}(\mu = 0, \sigma^2 = 1)$; thus, when v becomes larger, the $t(v)$ PDF and the Gaussian$(\mu = 0, \sigma^2 = 1)$ PDF look more alike.
- When compared to the standard Gaussian PDF, the t PDF, in general, is less peaked and has more probability (area) in the tails.

Definition 8.5. (CLT 3: Model Agnostic Unknown σ) Suppose that $X_1, X_2, ..., X_n$ is a random sample. Recall that Theorem 8.2 says the sample mean $\bar{X}$ has the following sampling distribution:

$$\bar{X} \sim \mathcal{AG}\left(\mu_{\bar{x}} = \mu_x, \sigma^2_{\bar{X}} = \frac{\sigma^2_x}{n}\right).$$

If we standardize $\bar{X}$, we obtain

$$Z = \frac{\bar{X} - \mu}{\sigma_x/\sqrt{n}} \sim \mathcal{AG}(\mu_z = 0, \sigma_z = 1).$$

If we replace the population standard deviation σ_x with the sample standard deviation S_x, we get a new sampling distribution:

$$t = \frac{\bar{X} - \mu_x}{s_x/\sqrt{n}} \sim \mathcal{AT}(n-1).$$

an approximate t distribution with degrees of freedom $v = n-1$. This means that the sampling distribution result approximately holds, even if the underlying population distribution is not perfectly Gaussian. The approximation is best when

- the sample size is larger
- the population distribution is more symmetric (not highly skewed).

Remark: Because normality (for the population distribution) is not absolutely critical for the t sampling distribution, we say that this sampling distribution is robust to the Gaussian assumption. Robustness is a nice property. Here, it assures us that the underlying assumption of normality is not an absolute requirement. Other sampling distribution results are not always robust to departures from normality.

Q: Why is the t distribution important?
A: It is often the case that the population standard deviation, σ_x, is unknown. This leaves us without a key value necessary for invoking Central Limit Theorem; e.g., we cannot use the following tools for large n.

$$Z_n = \frac{\bar{X} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{X} - \mu_x}{\sigma_x/\sqrt{n}} \sim \text{Gaussian}(0, 1)$$

$$\bar{X} \sim \mathcal{AG}(\mu_{\bar{x}} = \mu_x, \sigma_{\bar{x}} = \sigma_x/\sqrt{n}).$$

Under the same Central Limit Theorem assumptions, we have

$$T_n = \frac{\bar{X} - \mu_x}{s_x/\sqrt{n}} \sim \mathcal{AT} \text{ with } v \text{ degrees of freedom, } v = n - 1.$$

The difference between Z_n and T_n is between σ_x and s_X. As the sample size increases, this difference becomes negligible.

Summary: We will use the t statistic when working with quantitative data where σ_x is unknown (almost always).

Example 8.6. Hollow pipes are to be used in an electrical wiring project. In testing "1-inch" pipes, the data below were collected by a design engineer. The data are measurements of X, the outside diameter of this type of pipe (measured in inches). These $n = 25$ pipes were randomly selected and measured-all in the same location.

1.296	1.320	1.311	1.298	1.315
1.305	1.278	1.294	1.311	1.290
1.284	1.287	1.289	1.292	1.301
1.298	1.287	1.302	1.304	1.301
1.313	1.315	1.306	1.289	1.291

The manufacturers of this pipe claim that the population distribution is Gaussian (Normal) and that the mean outside diameter is $\mu_x = 1.29$ inches. Under this assumption (which may or may not be true), calculate the value of

$$t = \frac{\bar{x} - \mu_x}{s_x/\sqrt{n}}.$$

We use R to find the sample mean $\bar{x}$ and the sample standard deviation s_x:

```
> mean(pipes) ## sample mean
[1] 1.29908
> sd(pipes) ## sample standard deviation
[1] 0.01108272
```

We compute

$$t = \frac{1.299 - 1.29}{0.011/\sqrt{25}} \approx 4.096.$$

We plot the t distribution with $v = 25 - 1$ degrees of freedom in Figure 8.3.9, created with the following R code.

```
> ggdat<-data.frame(x=seq(-4.5,4.5,0.001),
+                   f=dt(x=seq(-4.5,4.5,0.001),df=25-1))
> ggdat.highlight<-data.frame(x=4.096,
+                             y=dt(x=4.096,df=25-1))
> ggplot(data=ggdat,aes(x=x,y=f))+
+   geom_line()+
+   geom_point(data=ggdat.highlight, aes(x=x,y=y), color="red")+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab(bquote(T))+
+   ylab(bquote(f[T](t)))+
+   ggtitle("T PDF")+
+   scale_color_discrete("")
```

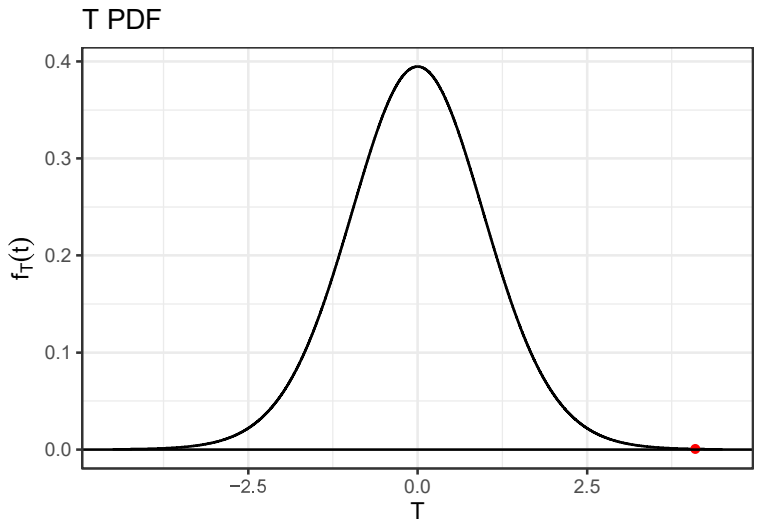

Figure 8.3.9: The $t(24)$. A red point at $t = 4.096$ has been added.

If the manufacturer's claim is true (that is, if $\mu_x = 1.29$ inches), then

$$t = \frac{\bar{x} - \mu_x}{s_x/\sqrt{n}}$$

should come from a t distribution with $v = 25 - 1$ degrees of freedom. We plotted the $t(24)$ PDF with a point at the value $t = 4.096$ in Figure 8.3.9.

Curiosity: Does $t = 4.096$ seem like a value you would expect to see from this distribution? Recall that t was computed under the assumption that $\mu_x = 1.29$ inches (the manufacturer's claim). Therefore, if the manufacturer's claim is true, we would expect to see a value of t around the center of this distribution. This isn't what we see here. This value of t is somewhat extreme (way out in the tail) and ultimately looks to be more consistent with a value of μ_x that is larger than 1.29 inches.

8.3.3 CLT and the Sampling Distribution of the Sample Proportion

Q: Can we apply this to the sampling distribution of the sample proportion?
A: Quickly, yes. The reason the Central Limit Theorem applies to the sample proportion is that we can view it as the success probability from a Bernoulli trial.

Consider a May 6-10, 2015 survey by Gallup that asked 1,024 American adults whether or not they consider married men and women having an affair is morally acceptable. Nine hundred and forty two respondents reported that the consider such behavior morally unacceptable leaving eighty two who consider such behavior morally acceptable.

Usually, we would calculate the sample proportion by taking the number with the attribute of interest and divide by the total number in the sample.

$$\widehat{p} = \frac{82}{1024} = 0.0801$$

Consider keeping track of all the responses using Bernoulli trials, e.g. recording ones for American

adults that consider married men and women having an affair morally acceptable and zeros for American adults who don't.

Q: What would the mean of these values be?
A: For this we have data that looks similar to this,

$$\begin{aligned} x_1 &= 1 \\ x_2 &= 0 \\ x_3 &= 0 \\ x_4 &= 0 \\ x_5 &= 1 \\ &\vdots \\ x_{1024} &= 0. \end{aligned}$$

The sum of this data,

$$\sum_{i=1}^{1024} x_i = 82,$$

is equal to the number of participants that consider married men and women having an affair morally acceptable. In this case, we can think of the sample proportion as the sample mean of Bernoulli observations.

$$\widehat{p} = \bar{x} = \frac{\sum_{i=1}^{1024} x_i}{1024}.$$

Definition 8.7. (CLT 4: Bernoulli Data) Suppose that $X_1, X_2, \ldots X_n$ is a random sample of Bernoulli observations with parameter p. As our sample size increases,

$$\widehat{P} \sim \mathcal{AG}\left(\mu_{\widehat{p}} = p, \sigma_{\widehat{p}} = \sqrt{\frac{p(1-p)}{n}}\right).$$

which implies that

$$Z_n = \frac{\widehat{p} - \mu_{\widehat{p}}}{\sigma_{\widehat{p}}} = \frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{AG}(\mu_z = 0, \sigma_z = 1)$$

We say that the distribution of Z_n is approximately the standard Gaussian distribution, the Gaussian distribution with mean zero and standard deviation one; and the distribution of $\widehat{P}$ is approximately Gaussian with mean $\mu_{\widehat{P}}$ and standard deviation $\sigma_{\widehat{P}}$.

Recall from Chapter 5, that $E(X) = p$ and $var(X) = p(1-p)$ for Bernoulli random variable X. We can take the parameters that describe the sampling distribution as

$$\begin{aligned} \mu_{\widehat{p}} &= E(X) = p \\ \sigma_{\widehat{p}} &= \sqrt{\frac{p(1-p)}{n}}, \end{aligned}$$

which shows this is a direct application of Theorem 8.2.

When talking about the sampling distribution of the sample proportion we say that the sampling distribution is approximately Gaussian if both of the following are true

- $np \geq 15$

- $n(1-p) \geq 15$.

note that np and $n(1-p)$ are the number of successes and failures in the sample, respectively. This is an more restrictive requirement than $n \geq 30$ in the previous cases.

It is important to note that this is just a guideline that works well for proportions. Another such guideline is to check $np(1-p) \geq 10$; which benchmark someone uses largely depends on the text they learned from. Figure **??** shows PMFs for the sample proportion varying n and p. As an exercise try checking the benchmarks for each case and assess it graphically.

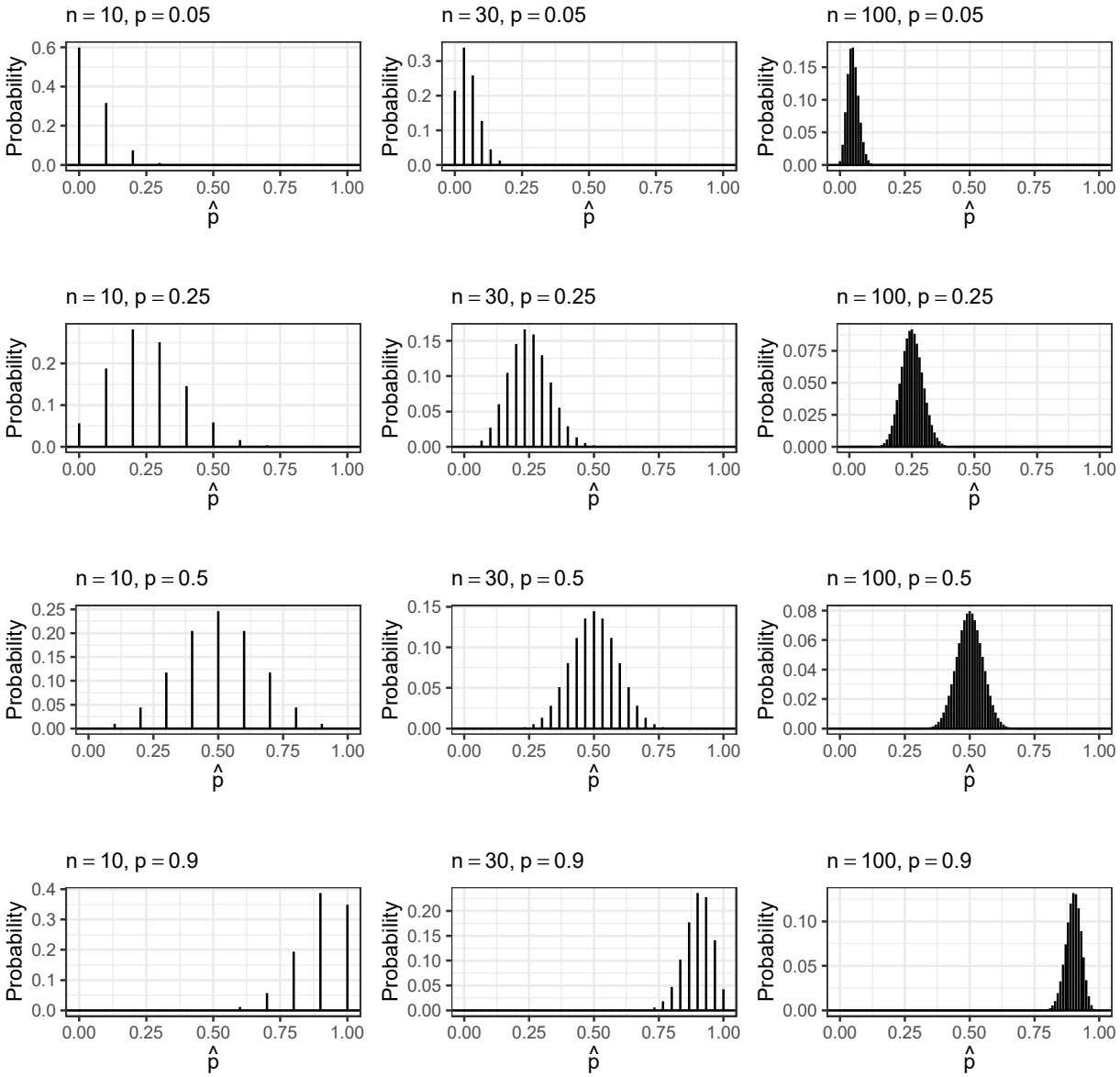

Figure 8.3.10: PMFs for the sample proportion varying n and p

Example 8.8. According to a recent Gallup report, just 16% of American adults approve of Congress (Gallup, 2017).

Q: What is the probability that, in a sample of one hundred American adults, twenty approve of

Congress?

A: Let X be the number of American adults that approve of Congress in a binomial experiment with $n = 100$ trials. We can calculate this probability using the Binomial PMF.

$$P(X = 20) = \binom{100}{20}(0.16)^{20}(1 - 0.16)^{80} = 0.0567$$

We can ask for this probability in R as follows.

```
> n=100
> p=0.16
> dbinom(x=20,size=n,prob=p)
[1] 0.05674059
```

This probability is the height of the bar at $X = 5$ in the population distribution of X depicted in Figure 8.3.11, which is created with the following R code.

```
> ggdat<-data.frame(x=0:n,
+                   f=dbinom(x=0:n,size=n,prob=p))
> ggdat.highlight<-data.frame(x=20,
+                             y=dbinom(x=20,size=n,prob=p))
> ggplot(data=ggdat,aes(x=x))+
+   geom_linerange(aes(ymax=f), ymin=0) +
+   geom_linerange(data=ggdat.highlight,aes(ymax=y), ymin=0,color="red") +
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab(bquote(X))+
+   ylab(bquote(f[X](x)))+
+   ggtitle("Binomial PMF",
+         subtitle=bquote(P(X==20)~"for"~n==100*","~p==0.16))
```

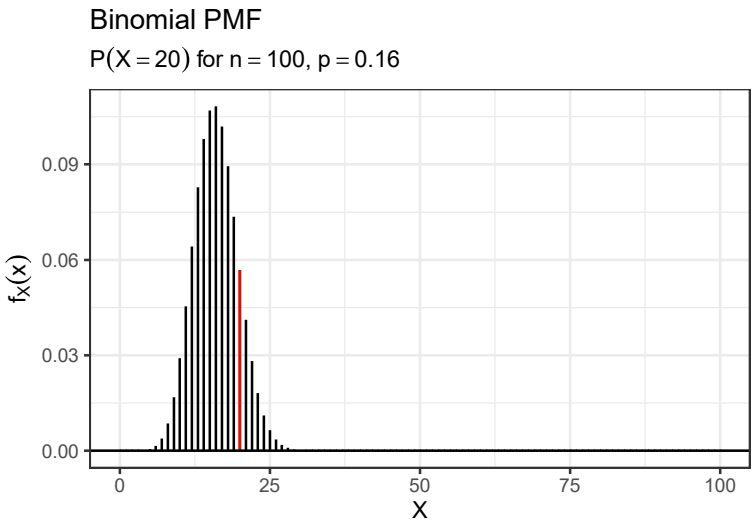

Figure 8.3.11: Binomial PMF for $n = 100$ and $p = 0.16$ with the bar at $X = 20$ highlighted in red which represents $P(X = 20)$ as calculated in Example 8.8.

Q: What is the approximate sampling distribution of the sample proportion from a sample of $n = 100$?

A: We can find the approximate sampling distribution of the sample proportion using the formulas

$$\mu_{\widehat{p}} = E(X) = p = 0.16$$
$$\sigma_{\widehat{p}} = sd(X) = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.16(1-0.16)}{100}} = 0.0367.$$

Although we can calculate probabilities about this experiment by viewing it as a Binomial experiment, we can also use the approximation of the sampling distribution of the sample proportion which we know to be approximately Gaussian as

$$np = 100(0.16) = 16 \geq 15$$
$$n(1-p) = 100(1-0.16) = 84 \geq 15.$$

Thus,

$$\hat{p} \sim \mathcal{AG}(\mu_{\widehat{p}} = 0.16, \sigma_{\widehat{p}} = 0.0367).$$

Q: What is the probability that more than twenty two of the random sample of 100 American adults approve of Congress?

A: Viewing this as a binomial experiment we get an exact probability by answering this question using the binomial CDF.

$$\begin{aligned} P(X > 22) &= 1 - (P(X = 0) + P(X = 1) + \ldots + P(X = 22)) \\ &= 1 - P(X \leq 22) \\ &= 0.0428 \end{aligned}$$

We can ask R for this probability as follows.

```
> n=100
> p=0.16
> 1-pbinom(q=22,size=n,prob=p)
[1] 0.04281426
```

This probability is the combined height of the bars at $X = 22$, $X = 23$, ..., $X = 100$ in the population distribution of X depicted in Figure 8.3.12, which is created with the following R code.

```
> ggdat<-data.frame(x=0:n,
+                   f=dbinom(x=0:n,size=n,prob=p),
+                   Fx=pbinom(q=0:n,size=n,prob=p))
> ggdat.highlight<-data.frame(x=23:n,
+                             y=dbinom(x=23:n,size=n,prob=p))
> g1<-ggplot(data=ggdat,aes(x=x))+
+   geom_linerange(aes(ymax=f), ymin=0) +
+   geom_linerange(data=ggdat.highlight,aes(ymax=y),ymin=0,color="red") +
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab(bquote(X))+
+   ylab(bquote(f[X](x)))+
```

```
+   ggtitle("Binomial PMF",
+         subtitle=bquote(P(X==22)~"for"~n==100*","~p==0.16))
> ggdat.openpoints<-data.frame(x=ggdat$x,
+                              y=pbinom(ggdat$x-1,size=n,prob=p))
> ggdat.closedpoints<-data.frame(x=ggdat$x,
+                                y=pbinom(ggdat$x,size=n,prob=p))
> ggdat.highlight<-data.frame(x=22,
+                             y=pbinom(q=22,size=n,prob=p))
> g1.CDF<-ggplot(data=ggdat, aes(x=x,y=Fx)) +
+   geom_step()+
+   geom_point(data = ggdat.openpoints, aes(x=x,y=y),shape=1) +
+   geom_point(data = ggdat.closedpoints, aes(x=x,y=y)) +
+   geom_point(data = ggdat.highlight, aes(x=x,y=y),color="red") +
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab(bquote(X))+
+   ylab(bquote(F[x](x)))+
+   ggtitle("Binomial CDF",
+         subtitle=bquote(P(X<=22)~"for"~n==100*","~p==0.16))
> grid.arrange(g1,g1.CDF,ncol=2)
[1] 0.04281426
```

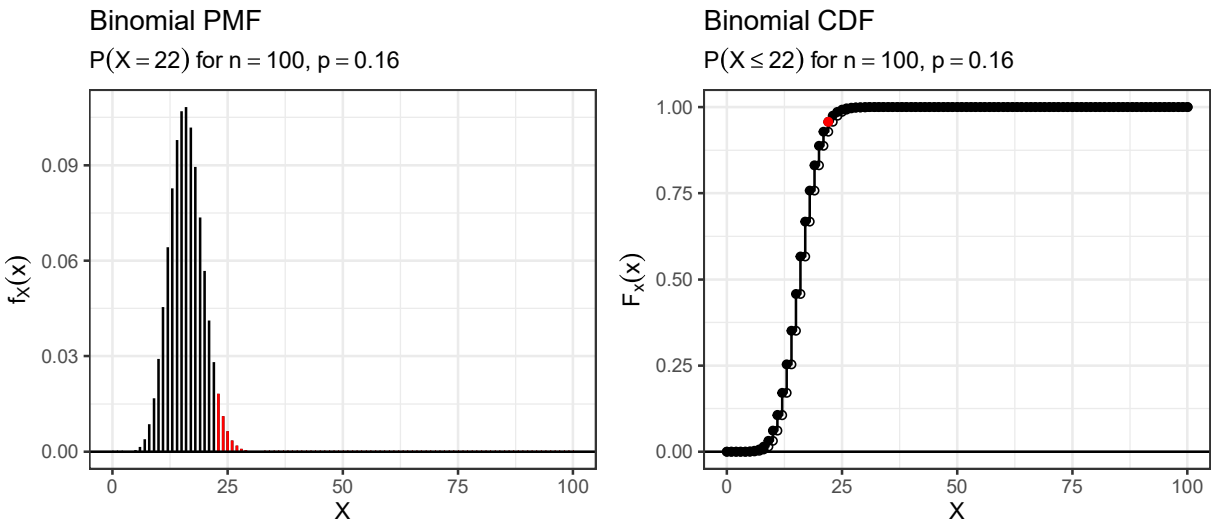

Figure 8.3.12: Binomial PMF for $n = 100$ and $p = 0.16$ with the bars at $X = 22$, $X = 23$, ..., $X = 100$ highlighted in red which represents $P(X > 22)$ as calculated in Example 8.8.

Q: What is the approximate probability that more than twenty two of the random sample of 100 American adults approve of Congress? **A:** We can also answer this question using the approximation of the sampling distribution of the sample proportion. Noting that $X = 22 \Rightarrow \hat{p} = \frac{22}{100} = 0.22$

$$\begin{aligned} P(\widehat{p} > 0.22) &= 1 - P(\widehat{p} \leq 0.22) \\ &= 0.0510 \end{aligned}$$

We can calculate this probability in R as follows.

```
> p<-0.16
> n<-100
> mu.p<-p
> sigma.p<-sqrt(p*(1-p)/n)
> 1-pnorm(q=0.22,mean=mu.p,sd=sigma.p)
[1] 0.05085347
```

This is the area under the sampling distribution as depicted in Figure 8.3.13, which is created with the following R code.

```
> ggdat<-data.frame(x=seq(0,0.35,0.001),
+                   f1=dnorm(x=seq(0,0.35,0.001),mean=mu.p,sd=sigma.p))
> g1<-ggplot(data=ggdat,aes(x=x))+
+   geom_line(aes(y=f1))+
+   geom_ribbon(data=subset(ggdat,x>0.22),aes(ymax=f1),ymin=0,
+               fill="red",colour=NA,alpha=0.5)+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab(bquote("Proportion that Approve of Congress"~(hat(p))))+
+   ylab(bquote(f[hat(p)](hat(p))))+
+   ggtitle("Gaussian PDF",
+        subtitle=bquote(P(hat(p)>0.22)~"for"~mu==0.16*","~sigma==0.0367))
> ggdat<-data.frame(x=seq(0,0.35,0.001),
+                   F1=pnorm(q=seq(0,0.35,0.001),mean=mu.p,sd=sigma.p))
> ggdat.highlight<-data.frame(x=0.22,
+                             y=pnorm(q=0.22,mean=0.16,sd=0.0367))
> g1.cdf<-ggplot(data=ggdat,aes(x=x))+
+   geom_line(aes(y=F1))+
+   geom_point(data=ggdat.highlight,aes(x=x,y=y),color="red")+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab(bquote("Proportion that Approve of Congress"~(hat(p))))+
+   ylab(bquote(F[hat(p)](hat(p))))+
+   ggtitle("Gaussian CDF",
+        subtitle=bquote(P(X<=40)~"for"~mu==0.16*","~sigma==0.0367))
> grid.arrange(g1,g1.cdf,ncol=2)
```

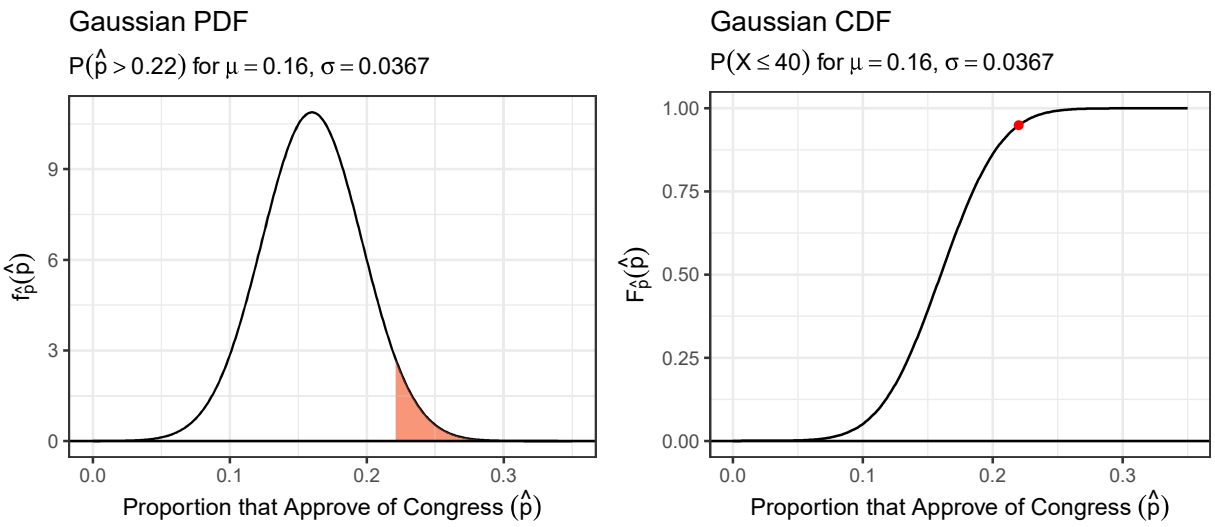

Figure 8.3.13: The sampling distribution, Gaussian PDF (left) for $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$ with area under the PDF shaded to the right of 22 representing $P(\bar{X} > 22)$ and CDF (right) with $P(\bar{X} \leq 0.22)$ highlighted with a red point.

This can be calculated and visualized in `R` as follows.

8.4 When Central Limit Theorem Doesn't Apply

As we saw in the previous sections, we can only apply the Central Limit Theorem to certain statistics, and when predefined benchmarks are met. What happens when the Central Limit Theorem doesn't apply? How can we assess the sampling distributions of $\widehat{P}$ and $\bar{X}$? How can we assess the sampling distribution of other statistics? The answer to these questions is resampling.

Definition 8.9. Resampling is a technique that involves taking repeated random samples from an original, representative sample to mimic reproducing an experiment on the population. For each random sample taken from the original, we can calculate summary statistics like $\widehat{p}$, $\bar{x}$.

To assess the sampling distribution of statistics like $\widehat{p}$ or $\bar{x}$ when the assumptions for using the Central Limit Theorem aren't met we can perform resampling, plot a histogram of the statistics for all resamples and a **sampling density curve**. This curve estimates the sampling distribution.

Example 8.10. Consider clinical trials conducted at Stanford University to evaluate the efficacy of maintenance chemotherapy for acute myelogenous leukaemia (Embury et al., 1977). After reaching remission, patients were randomized into two groups. One group of patients in this study received maintenance chemotherapy even though they were cancer-free. Below are the data for the time in weeks until relapse of the patients in the maintenance chemotherapy group.

9, 13, 13, 18, 23, 28, 31, 34, 45, 48, 161

A histogram of this data with an estimate of the population density curve can be seen in Figure 8.4.14 and is created using the following `R` code.

```
> dat<-data.frame(remission.time=c(9, 13, 13, 18, 23, 28, 31, 34, 45, 48, 161))
> ggplot(data=dat,aes(x=remission.time))+
+   geom_histogram(aes(y=..density..),breaks=seq(5,170,20),fill="lightblue",color="black")+
+   geom_density(color="red")+
+   theme_bw()+
+   xlab("Weeks until Relapse (X)")+
+   ylab("Density")+
+   ggtitle("")
```

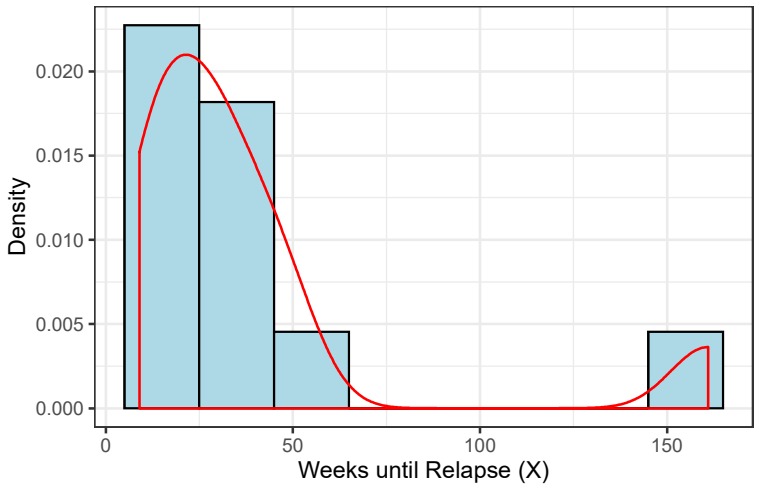

Figure 8.4.14: A histogram of the data from Example 8.10 with an estimate of the population density curve in red.

Here, the data shows it is unlikely that the population distribution is Gaussian and the sample size is too small to meet our benchmark for using Central Limit Theorem.
Q: What is the sampling distribution of $\bar{X}$?
A: Without the ability to simply apply the Central Limit Theorem, this question requires advanced probability training to reveal the true PDF for the sampling distribution. For now, until you take a full semester of probability, we will lean on resampling techniques to give us a good approximation of the desired sampling distribution.

Assuming that our sample is representative of the population, we can sample from it as if it were the population and assess the values we get back.

```
> remission.time=c(9, 13, 13, 18, 23, 28, 31, 34, 45, 48, 161)
> samp<-sample(x=remission.time,size=11,replace=TRUE)
> mean(samp)
[1] 41.54545
> samp<-sample(x=remission.time,size=11,replace=TRUE)
> mean(samp)
[1] 61.72727
> samp<-sample(x=remission.time,size=11,replace=TRUE)
> mean(samp)
[1] 38.81818
```

```
> samp<-sample(x=remission.time,size=11,replace=TRUE)
> mean(samp)
[1] 48.09091
```

We see that for each sample we yield a different sample mean. This, in a sense, simulates doing the experiment over and over again. Figure 8.4.15, shows a histogram of 1,000 of these sample means from resampling with an estimate of the sampling density curve in red.

```
> means<-rep(NA,1000)
> for(i in 1:1000){
+   samp<-sample(x=remission.time,size=11,replace=TRUE)
+   means[i]<-mean(samp)
> }
> dat<-data.frame(means=means)
> ggplot(data=dat,aes(x=means))+
+   geom_histogram(aes(y=..density..),bins=15,fill="lightblue",color="black")+
+   geom_density(color="red")+
+   theme_bw()+
+   xlab(bquote("Average Weeks until Relapse"~(bar(X))))+
+   ylab("Density")+
+   ggtitle("")
```

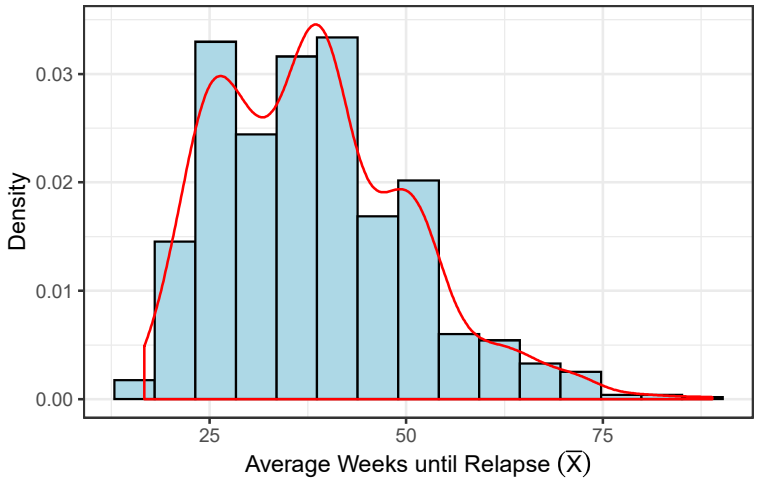

Figure 8.4.15: A histogram of 1,000 sample means from Example 8.10 with an estimate of the sampling density curve in red.

We see that the sampling distribution for $\bar{X}$ is right skewed, and thus not Gaussian; this is expected as the benchmarks for applying Central Limit Theorem aren't met. If we had blindly applied Central Limit Theorem here, our assessments based on that assumption would be incorrect.

Example 8.11. Croissant (2016) make available data about reported total labor incomes from $n = 4,586$ Americans in 1993. This data is plotted in Figure 8.4.16.

```
> library(Ecdat)
> incomes<-PSID$earnings
> ggdat<-data.frame(incomes=incomes)
> ggplot(data=ggdat,aes(x=incomes))+
+   geom_histogram(aes(y=..density..),bins=30,fill="lightblue",color="black")+
+   geom_density(color="red")+
+   theme_bw()+
+   xlab("Total Labor Income (X)")+
+   ylab("Density")+
+   ggtitle("")
```

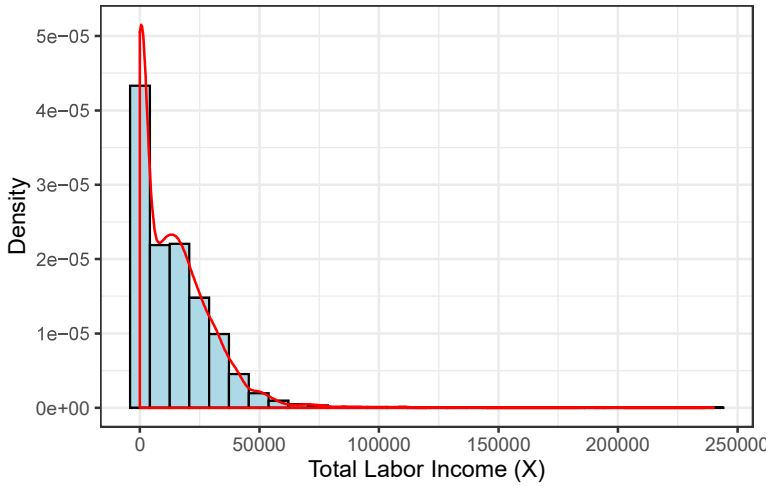

Figure 8.4.16: A histogram of the total labor incomes reported by the $n = 4,586$ Americans.

We know that, since $n > 30$, the Central Limit Theorem assumptions are met and that the sampling distribution for $\bar{X}$ will be approximately Gaussian. This can be confirmed by plotting the means yielded from resampling as seen in Figure 8.4.17.

```
> means<-rep(NA,1000)
> for(i in 1:1000){
+   samp<-sample(x=incomes,size=4856,replace=TRUE)
+   means[i]<-mean(samp)
+ }
> dat<-data.frame(means=means)
> ggplot(data=dat,aes(x=means))+
+   geom_histogram(aes(y=..density..),bins=15,fill="lightblue",color="black")+
+   geom_density(color="red")+
+   theme_bw()+
+   xlab(bquote("Average Labor Income"~(bar(X))))+
+   ylab("Density")+
+   ggtitle("")
```

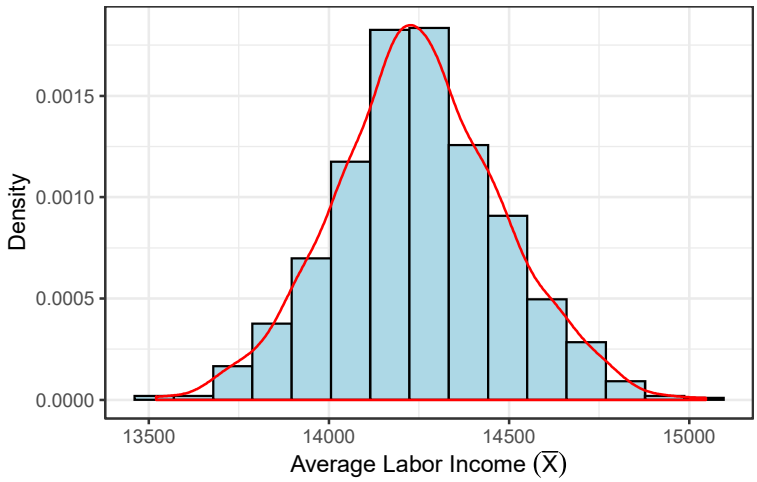

Figure 8.4.17: A histogram of 1,000 sample means from Example 8.11 with an estimate of the sampling density curve in red.

Q: Can we use resampling to estimate the sampling distribution for other statistics like the sample median ($\widehat{m}$)?
A: Yes, and for this example it is a very important question. Figure 8.4.16, shows a histogram of the total labor incomes reported. We know that the sample median is a more responsible measure of center for such data and so we might want to ask questions about that.

The Central Limit Theorem does not apply to finding the sampling distribution of $\widehat{m}$, and so we can use resampling to assess the sampling distribution of $\widehat{m}$ (until we take a probability course).

We can estimate the sampling distribution of $\widehat{m}$ by plotting the sample medians yielded from resampling, seen in Figure 8.4.18. We see that the sampling distribution of $\widehat{m}$ is not Gaussian and appears to be multi-modal. While we don't have a name for this sampling distribution, our ability to estimate it will have much importance when making inference about the population median in the following chapters.

```
> median<-rep(NA,1000)
> for(i in 1:1000){
+   samp<-sample(x=incomes,size=4856,replace=TRUE)
+   median[i]<-median(samp)
+ }
> dat<-data.frame(median=median)
> ggplot(data=dat,aes(x=median))+
+   geom_histogram(aes(y=..density..),bins=15,fill="lightblue",color="black")+
+   geom_density(color="red")+
+   theme_bw()+
+   xlab(bquote("Median Labor Income"~(hat(m))))+
+   ylab("Density")+
+   ggtitle("")
```

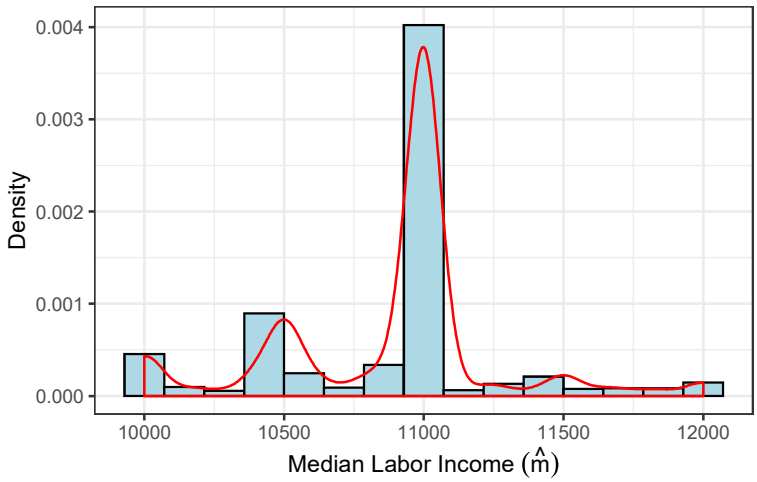

Figure 8.4.18: A histogram of 1,000 sample medians from Example 8.11 with an estimate of the sampling density curve in red.