# Contents



# Part I

# Statistics

**Model**

Now that you know some probability, you are ready to begin building the models useful for analyzing data.

**Why do we need a *model*?** We will not progress far without a model, *some collection of hypotheses*, about how the population works. Suppose you treat two sets of patients for high blood pressure, one with treatment $A$ and the other with treatment $B$. Your goal is to determine whether there is a difference in the two treatments. Unfortunately, no two patients respond in exactly the same way. Some report getting better, some report getting worse. To make some kind of sense of the mess, you decide to measure the patients' blood pressure one week after treatment starts, but the picture is still not clear. On both treatments, some blood pressures go down, some go up. You then decide to average the blood pressure across patients, obtaining an average change under treatment $A$ and another under treatment $B$. Finally the data start to make sense. Treatment $A$ produces a greater average reduction in blood pressure than treatment $B$. But, you are savvy enough to realize that there was so much variation among your patients that the average numbers might be reversed if you were to do the experiment again. How can you be sure that the difference is substantial? How can you be sure that treatment $A$ is actually better than treatment $B$?

You need to understand how average blood pressure changes *vary*. One way is to repeat the study multiple times, but it is usually not a feasible solution. If instead you had a good *model* of the variation, you would be able to predict the outcome of future experiments without doing them! Then, if all or most of the predictions showed treatment $A$ performing better than treatment $B$, you could be fairly convinced that treatment $A$ was in fact better than treatment $B$.

**How are models constructed?** Models are constructed from common sense, for convenience, from theory, and sometimes from untruths that you expect will be disproven in the face of data. Models are *always* wrong,

but that does not make them useless. Even wrong models may be accurate for their purpose. If they produce the right conclusion most of the time they can be exceptionally valuable. Mathematical results coming out of the theoretical side of statistics have produced some startling simple, but useful models that apply in a wide range of situations. These theoretical models will help you construct models for your applications. However, one *must* understand the limits of these theoretical results. Because *all* models are wrong, you must know when they are wrong and insure they are not wrong under the conditions you are using them. *Blind use of any model is dangerous.*

Models in statistics are *probability models*, meaning they produce a random outcome given a set of inputs. As we now know, random does not mean these models lack any information. Probability models are essential in statistics because they account for

- the inevitable measurement errors associated with data collection,
- some of the uncertainty due to the fact that every model is only an approximation to reality, and
- the variation caused by sampling only a small portion of the entire population.

**Goals of Statistics**

With probability models in hand, the goals of statistical analyses can be categorized into three main areas:

- Estimates population properties.
- Evaluate plausibility of current hypotheses about the population.
- Predict future observations from the population.

**Population**

The *population* is a key concept. When we collect data, we do so in order to learn about a population. For example, we survey a handful of voters to learn how the entire voting population will act in the upcoming election. Or we test a subset of toys coming off the assembly line to determine whether the entire population of toys shipping from a factory meet certain quality requirements.

The first lesson to learn about populations, is that you should never generalize your conclusions to populations from which you have sampled no data.

Or stated in another way, be very careful how you sample your data to insure that the conclusions you draw will apply to the population that interests you.

**Random Sample**

Key to making inference about a population is the sample that you take from it. A *random sample* taken from a population has very specific properties that insure conclusions draw from its analysis will extend back to the population.

Definition: *random sample*

> A *random sample* is a collection of individuals, where individuals are chosen
>
> - from the same population, and
> - independently of each other.

If the population size $N$ is small, and the sample size $n$ is large, then the independence assumption is not reasonable because who we sample next, depends on who we sampled before. However, we typically work under conditions where $n << N$, so that sampling can be *effectively independent*. It is still up to the experimenter to insure that the sampling strategy satisfies the above two properties.

# 1 Univariate Data

## 1.1 Sampling Distribution

**Sampling Distribution of $X$**

Suppose we measure some variable on a *random sample* of $n$ individuals drawn from a *population*, producing measurements $X_1, X_2, \ldots, X_n$. These are random variables, which under the assumptions of a random sample, follow some distribution

$$X_i \overset{\text{iid}}{\sim} F(x;\theta)$$

for all $i = 1, \ldots, n$. Here, $F(x;\theta)$ is some cdf (remember we said the cdf of a random variable always exists) that depends on parameters $\theta$.

Definition: *sampling distribution of $X$*

> The distribution $F(x,\theta)$ is called the *sampling distribution* of the data.

We need the sampling distribution in order to interpret the randomness in the data. How do we come up with a sampling distribution?

**Proposing a Sampling Distribution**

In general, it is a very difficult task. We will often get around this kind of challenge by looking at summary measures of the data, rather than the raw data. This class will provide you with a toolbox of useful sampling distributions for data summaries that will be related to the "named" distributions you learned in probability theory. These toolbox distributions may be all you ever need. In fact, very often, we may manipulate the data until it satisfies the assumptions of a well-known sampling distribution just to avoid the heavy task of deriving our own sampling distribution.

Nevertheless, you should always examine your raw data and think about its properties. Big, blatant errors are amazingly common, and no amount of good statistics is going to save your from the wrong conclusion when errors go uncorrected.

Here are two approaches to help you "look" at your data.

- **Numerical summaries.** Compare numerical summaries of your data with what you expect.
- **Graphical summaries.** Compare graphical summaries of your data with what you expect.

### 1.1.1 Numerical Summaries

**Statistic**

Definition: *statistic*

> A *statistic* is any function $t(X)$ of the data $X = (X_1, \ldots, X_n)$.

Since a *statistic* is a function of random variables, it is itself a random variable with its own *sampling distribution*. There is a sampling distribution for every possible statistic of your data, although very few of them are known and named.

Sample statistics $t(X)$ are to be clearly distinguished from population parameters $\theta$. Statistics are random variables. Parameters are not. The value of a statistic can be computed given data. The value of the population parameter can never be known unless you can sample the entire population. Usually, however, it can only be *estimated*.

To illustrate the various statistics we will define, we will use a dataset from the class and `R`.

```
> d <- read.csv("stat430.csv", header=T)
> head(d[,c(18,20,21,35:37)])
FirstName   LastName Major RExam Sheets Typed
1 Arunkumar Asaithambi   BMS    13    0.0     N
2 Sahnghyun        Cha  COMS    40    2.5     Y
3     Ruchi  Chaudhary  COMS    NA     NA  <NA>
4    Yetian       Chen  COMS    76    8.0     N
5    Yinbin       Chen LOMIS    38    2.0     N
6       Hye        Cho   BCB    59    4.0     N
```

**Numerical Statistics**

- **sample mode.** The observation occurring most often in the sample. It is directly observable for discrete data. For continuous data, it can be estimated when the cdf $F(x;\theta)$ is known (e.g. normal distribution).

  ```
  > table(d$Sheets)

  0    1  1.5    2  2.5    3    4    5    6  6.5    8   11   13 31.5
  1    2    1    3    1    1    3    4    1    1    3    1    2    1
  ```

  The output shows the frequency of each number in the dataset. Notice, one person had 31.5 pages in their cheatsheet.

- **sample median.** The middle value in the data, when arranged lowest to highest. (Henceforth, I will not show `R` output unless something needs to be explained.)

  ```
  > median(d$Sheets, na.rm=T)
  ```

- **sample mean.** The balancing point of the dataset, if values $X_i$, each weighing the same amount, are arranged on the real line.

  $$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

  ```
  > m <- mean(d$Sheets, na.rm=T)
  ```

  The mean tends to be sensitive to *outliers* in the dataset, moreso than the median. Outliers, are data points that are unlike the rest. If you accidently sampled $X_3 \sim G(x;\theta)$ when the rest $X_1, X_2, X_4, \ldots, X_n \overset{\text{iid}}{\sim} F(x;\theta)$, then $X_3$ may look like an outlier if $G$ is very different from $F$.

  Examining the output of the `table()` command, we observe one obvious outlier. We now remove it from the dataset to see its effect on the mean.

  ```
  > d.trunc <- d[d$Sheets!=31.5 & !is.na(d$Sheets),]
  > m.trunc <- mean(d$Sheets, na.rm=T)
  ```

- **sample range.** The difference between the largest and smallest measurement in the data.

  ```
  > range(d$Sheets, na.rm=T)
  ```

- **sample quantile.** The $\phi_q$ such that $\frac{\#\{X_i \le \phi_q\}}{n} \ge q$.

  ```
  > quantile(d$Sheets, prob=0.5, na.rm=T)
  ```

  The above is also known as the median.

- **interquartile range.** $\phi_{0.75} - \phi_{0.25}$

```
> IRQ(d$Sheets, na.rm=T)
```

- **sample deviation.** Each measurement has a deviation $\Delta_i = X_i - \bar{X}$
- **sample variance.** A measure of the spread in the sample.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- **sample standard deviation.** A measure of the spread of a sample, on the same scale as the original measurements,

$$s = \sqrt{s^2}$$

- **Empirical Rule.** For bell-shaped curves, we can predict how many data points are expected to fall within certain ranges.

| | |
|---|---|
| $\bar{X} \pm s$ | 68% data points expected here |
| $\bar{X} \pm 2s$ | 95% data points expected here |
| $\bar{X} \pm 3s$ | 99.7% data points expected here |

```
> sum(d$Sheets>m-s & d$Seehts < m+2, na.rm=T)/sum(!is.na(d$Sheets)
```

- **coefficient of variation.** The coefficient of variation is

$$CV = \frac{s}{|\bar{X}|}$$

Example:

> Suppose the mean weight of fertilizer bags is $\bar{X} = 80.6$ and the standard deviation is $s = 1.2$, both measured in pounds. Supose the mean weight of cereal boxes is $\bar{X} = 24.308$ and the standard deviation $s = 0.4$, both measured in ounces. Which process of filling bags is more error prone?
>
> Since the measurements are taken on different scales, it is difficult to make a comparison. We can, however, compare sample CV's. The comparison
>
> $$0.01488834 = \frac{1.2}{80.6} < \frac{0.4}{24.308} = 0.01645549$$
>
> reveals that the cereal box filling process is slightly more error-prone.

### 1.1.2 Graphical Summaries

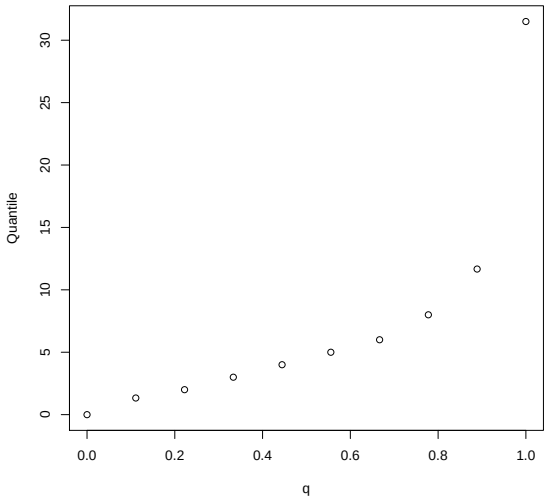
**Graphical Summaries**

- **Histogram.**

```
> hist(d$Sheets, freq=F, ylab="Frequency", xlab="No. Sheets")
```

- **Quantile Plot.** Plot probability $q$ against quantiles $\phi_q$ to obtain a "quantile plot."

```
> q <- seq(from=0, to=1, length.out=10)
> plot(x=q, y=quantile(d$Sheets, prob=q, na.rm=T), xlab="q", ylab="Quantile")
```



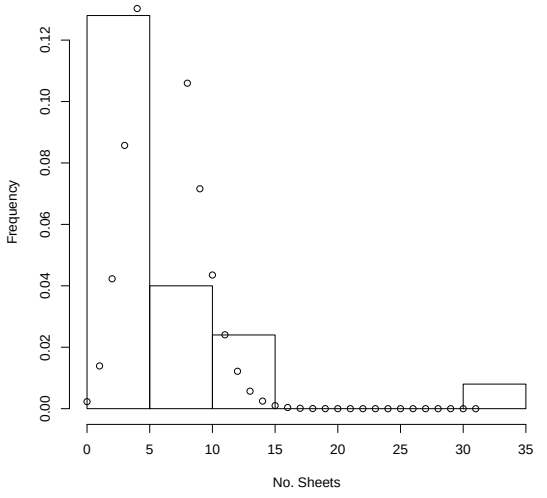

**Example: Cheat Sheet Distribution**

Sometimes, it still falls on our shoulders to propose a sampling distribution for the raw data. Let's make an attempt with the `d$Sheets` data (although we will not be very successful).

The data is more-or-less discrete, although I did record half-pages sometimes. If we round those numbers up, then all values are integers.

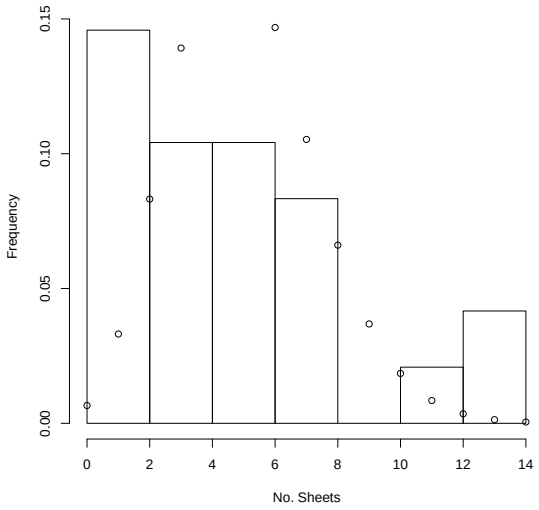
Among the discrete distributions, the Poisson distribution seems like a reasonable choice since there is no upper limit on the random variable $X$. The Poisson distribution has one parameter $\lambda$, which is its expectation. We will talk more about it shortly, but we can estimate the population mean by using the sample mean, thus $\hat{\lambda} = m$. Then, our proposed Poisson is fully specified. Let's compare it to the data.

```
> hist(d$Sheets, freq=F, ylab="Frequency", xlab="No. Sheets", main="")
> x <- min(d$Sheets, na.rm=T):max(d$Sheets, na.rm=T)
> points(x=x, y=dpois(x=x, lambda=m))
```




Not a great fit. Perhaps if we removed the outlier, things would be better...

```
> hist(d.trunc$Sheets, freq=F, ylab="Frequency", xlab="No. Sheets", main="")
> x <- min(d.trunc$Sheets, na.rm=T):max(d.trunc$Sheets, na.rm=T)
> points(x=x, y=dpois(x=x, lambda=m.trunc))
```




Not much better.

We could go on proposing distributions until we found a visual match. Later we will talk about how to quantify that match.

We could also consider numerical summaries.

```
> mean(d.trunc$Sheets)
[1] 5.020833
> var(d.trunc$Sheets)
[1] 13.20607
```

The mean and variance of a Poisson distribution should be identical. The large difference in sample mean and variance suggest again that a Poisson is a poor choice. Still, we have been very qualitative in our analysis. That is entirely appropriate for a preliminary analysis. Once a sampling distribution $F(x;\theta)$ has been proposed, we can become much more quantitative and decisive about our choices.

## 1.2 Central Limit Theorem

Earlier, I stated that we will do most of our statistics with numerical summaries of the data because theoretical sampling distributions exist for such summaries. We now begin several lectures on this theory. We won't go through all the gory details for all the derivations (just a few of the simpler ones). Your job is to make note of all the assumptions required to use these sampling distributions.

Our first result is for the *sample mean*.

**Central Limit Theorem**

**Theorem 1.** *Suppose* $X_1, \ldots, X_n \overset{iid}{\sim} F(x;\theta)$ *with mean* $\mu$ *and variance* $\sigma^2 < \infty$. *Then*

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

*or, equivalently*

$$n\bar{X} \sim N(n\mu, n\sigma^2)$$

One cannot really overestimate the power of this result. Just two of the immediate consequences are:

- The sampling distribution of $\bar{X}$ is available with very little additional information (just mean and variance of *data sampling distribution* $F(x;\theta)$).
- It allows us to approximate distributions when they become difficult to calculate, i.e. Binomial with large $n$.

We will see more consequences in the future.

**Corollary 2.** *There are some immediate corollaries that follow from the CLT.*

- *Binomial*$(n,p)$ *is asymptotically* $N(np, np(1-p))$ *when* $n$ *is large because* $X = X_1 + \cdots X_n$ *is a sum of iid Bernoulli.*
- *Poisson*$(\lambda)$ *is asymptotically like* $N(\lambda,\lambda)$ *because* $Y = Y_1 + \cdots Y_\lambda$ *is a sum of iid Poisson*$(1)$ *random variables. Yes,* $\lambda$ *need not be an integer, but it is growing large, so round it to the nearest integer when constructing the sum.*

**Examples**

Example:

> Suppose the number of hits on a web page in one hour is Poisson$(1000)$. What is the probability there will be more than $1,100$ hits in one hour?
>
> $X \sim$Poisson$(1000)$, so $P(X > 1100) =$ `1 - ppois(1100, lambda=1000)` or *approximately* `1-pnorm(1100, mean=1000, sd=sqrt(1000))`

```
> 1 - ppois(1100, lambda=1000)
[1] 0.000867641
> 1 - pnorm(1100, mean=1000, sd=sqrt(1000))
[1] 0.0007827011
```

Example:

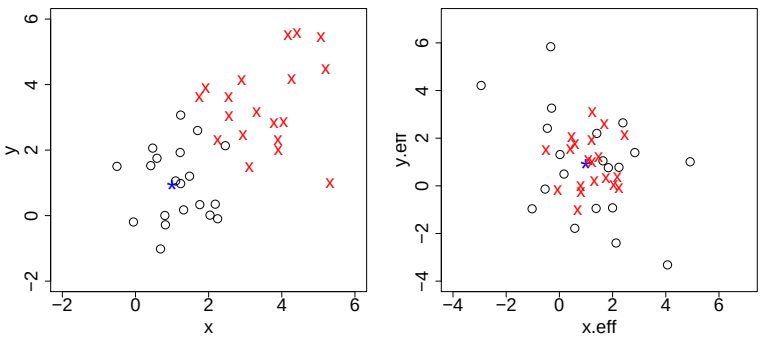
> Suppose $U_1, \ldots, U_5 \overset{\text{iid}}{\sim}$ Unif$(0,1)$. What is the distribution of $\bar{U}$?
>
> In general, we have no idea, but approximately
>
> $$\bar{U} \sim N(0.5, 1/60).$$

**Goals of Statistics**

We now restate the three main goals of statistics, adding some notation that is now in our vocabulary.

- Estimates population parameters $\theta$.
- Evaluate plausibility of proposed values of $\theta$ and current hypotheses $F(x;\theta)$ about the population.
- Predict future observations from the population.

We now tackle the first, parameter estimation.




(a) Bias (b) Efficiency

Figure 1: Desirable Properties of Estimators. True parameter is the symbol '*' in blue. Better estimator is black 'o', and worse estimator is red 'x'.

## 1.3 Parameter Estimation

### Parameter estimation

Definition: *estimator*

> Given $X_1, \ldots, X_n \sim F(x; \theta)$, then if $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ is used to estimate $\theta$, it is called an *estimator* of $\theta$. Notice, it is both a *statistic* and a *random variable*. Furthermore, the particular value $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$ come from evaluating the function given realization of the data $X = (x_1, \ldots, x_n)$, is called an *estimate* of $\theta$.

### Properties of Estimators

Much of statistics is about finding better estimators of population quantities. If we are presented with two estimators, how do we determine which is better? There are a few desirable properties that all good estimators should have.

- **Unbiased.** Estimator $\hat{\theta}$ is unbiased (Fig. 1(a)) if

$$E[\hat{\theta}] = \theta$$

- **Efficient.** Estimator $\hat{\theta}_1$ is more efficient (Fig. 1(b)) than estimator than $\hat{\theta}_2$ if

$$\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2)$$

- **Consistent.** An estimator is consistent if

$$\lim_{n \to \infty} P(|\hat{\theta} - \theta| > \epsilon) = 0$$

### Example: sample mean

The sample mean is the balancing center of a sample. It is logical to propose it as an estimator of the balancing center of a distribution, i.e. the distribution mean $\mu$. Furthermore, trivial calculations

$$E[\bar{X}] = E\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

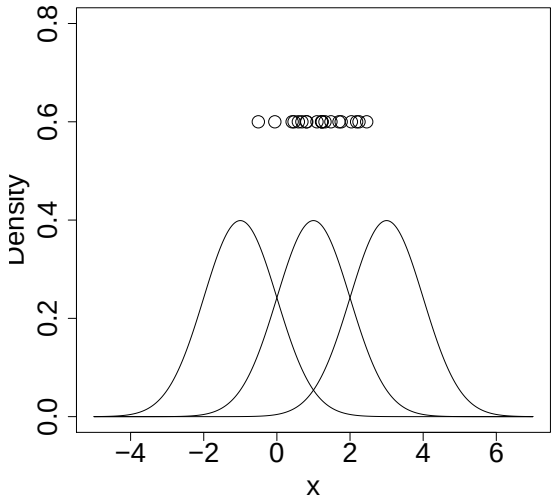



Figure 2: Maximum Likelihood Estimation

show that it is an unbiased estimator.

A little more work

$$\text{Var}(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}$$

show that its efficiency increases with sample size $n$. Most estimators have this pleasing property.

### 1.3.1 Maximum Likelihood Parameter Estimation

**Maximum Likelihood Method**

Suppose you collect some data $X_1, \ldots, X_n$. Example data are plotted in Fig. 2 along the line $y = 0.6$. We hypothesize that the data come from sampling distribution $F(x;\theta)$, but we don't know the value of the population parameter $\theta$. Below the data, we plot the probability density function $f(x)$, corresponding to cdf $F(x;\theta)$ for three different values of $\theta$. The peaks of the pdf indicate which values of $x$ are most likely realization of random variables $X_i$. We expect that a good value of $\theta$ is one that matches of the peaks in the pdf with the highest density areas of the data. Taking the best match, we select the middle density as most representative of the data.

The goal of the maximum likelihood method of estimation is to find $\theta$ that creates the best match between the theoretical pdf and the observed data $X$. We now describe how this matchup is done quantitatively.

**Likelihood Function**

Definition: *likelihood function*

> Given data $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F(x;\theta)$, the joint distribution of a realization $X = x$ of the data is
>
> $$f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i;\theta)$$
>
> where $f(x;\theta)$ is the pdf obtained from cdf $F(x;\theta)$. The likelihood function is simply this joint density
>
> $$L(\theta; x) = L(\theta) = f(x_1, \ldots, x_n; \theta)$$

viewed as a function of the unknown population parameter(s) $\theta$. Throughout, we have made the other dependences explicit after the semicolon ';', but we may drop these in the notation and make the dependence implicit in practice.

**Maximum Likelihood Estimator**

Definition: *maximum likelihood estimator* $\hat{\theta}$

The maximum likelihood estimator (MLE) is

$$\hat{\theta} = \text{argmax}_\theta L(\theta; X)$$

Equivalently, we may define the MLE as

$$\hat{\theta} = \text{argmax}_\theta \ln L(\theta; X)$$

the value of the parameter(s) $\theta$ that maximizes the *log likelihood* function, often written $l(\theta; x)$.

In practice, you will find it generally much easier to work with the log likelihood, especially when you compute likelihoods with computers. Consider the following code from `R`:

```
> log(1e-200)
[1] -460.517
> log(1e-200*1e-200)
[1] -Inf     # huh?
> 1e-200*1e-200
[1] 0        # ah!  computer thinks 1e-400 is 0
> log(1e-200)+log(1e-200)
[1] -921.034 # everything is all right if summed on the log scale
```

We see that when the likelihood drops to $1 \times 10^{-400}$, the computer considers the values 0. Logging it now cannot rescue us from round-off error. The likelihood is usually $\prod_{i=1}^n f(x_i)$, a product of small numbers, which is itself very small. On the log scale, the log likelihood is a sum $\sum_{i=1}^n \ln f(x_i)$ of negative numbers. Far more calculations can be completed without round-off error on the log scale.

**Example: Binomial Random Variable**

Example:

Suppose you observe $Y = y_0 \sim \text{Binomial}(n, p)$ where $n$ is the number of trials known beforehand, and $p$ is the unknown probability of "success." Use the maximum likelihood method to produce a maximum likelihood estimator of $p$.

$$\begin{aligned}
L(p; y_0) &= \tbinom{n}{y_0} p^{y_0} (1-p)^{n-y_0} && \text{likelihood} \\
l(p; y_0) &= \ln \tbinom{n}{y_0} + y_0 \ln p + (n - y_0) \ln(1-p) && \text{log likelihood} \\
\tfrac{dl}{dp} &= \tfrac{y_0}{p} - \tfrac{n-y_0}{1-p} && \text{take derivative to find max} \\
\tfrac{y_0}{\hat{p}} - \tfrac{n-y_0}{1-\hat{p}} &= 0 && \text{mle solves this eqn.} \\
y_0(1-\hat{p}) - (n-y_0)\hat{p} &= 0 && \text{algebra...} \\
\hat{p} &= \tfrac{y_0}{n} && \text{MLE} \\
\tfrac{d^2 l}{dp^2} &= -\tfrac{y_0}{p^2} - \tfrac{n-y_0}{(1-p)^2} < 0 && \text{Verify maximum}
\end{aligned}$$

**Example: Poisson Random Variable**

Example:

The number of red cars pulling into lot #22 between 8:30 and 8:40 am Monday through Friday are counted for several days to produce data

$$X = (3, 2, 3, 3, 4, 1, 4, 2, 4, 3)$$

What would be a reasonable sampling distribution for this data? Estimate its parameters.

We assume that the data from different days are independent and identically distributed. These are count data with realizations in space $\{0, 1, 2, \ldots\}$. If we assume red car arrivals are independent and rare at any particular instant of time, then this situation would seem to satisfy the Law of Rare events, which leads to the Poisson distribution.

$$\begin{aligned}
L(\lambda) &= \prod_{i=1}^{10} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} && \text{likelihood: independence and pmf of Poisson} \\
L(\lambda) &= e^{-10\lambda} \lambda^{\sum_{i=1}^{n} x_i} \prod_{i=1}^{10} \frac{1}{x_i!} && \text{rearrangement} \\
l(\lambda) &= -10\lambda + \ln\lambda \sum_{i=1}^{10} x_i - \sum_{i=1}^{10} \ln(x_i!) && \text{log likelihood} \\
\frac{dl}{d\lambda} &= -10 + \frac{\sum_{i=1}^{10} x_i}{\lambda} && \text{take derivative for max} \\
-10 + \frac{\sum_{i=1}^{10} x_i}{\hat{\lambda}} &= 0 && \text{MLE } \hat{\lambda} \text{ satisfies this eqn} \\
\hat{\lambda} &= \bar{X}
\end{aligned}$$

**Example: Normal Random Variable**

Example:

Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$. Find the MLEs $\hat{\mu}$ and $\hat{\sigma}^2$.

$$\begin{aligned}
L(\mu, \sigma^2) &= \prod_{i=1}^{n} f(x_i) && \text{where } f(x_i) \text{ is the pdf of normal} \\
&= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{(} x_i - \mu)^2\right] && \text{algebra} \\
l(\mu, \sigma^2) &= -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 && \text{log likelihood} \\
\frac{dl}{d\mu} &= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} -2(x_i - \mu) && \text{derivative wrt } \mu \\
\hat{\mu} &= \bar{X} && \text{sample mean, once again} \\
\frac{dl}{d\sigma^2} &= \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 && \text{derivative wrt } \sigma^2 \\
\frac{-n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 &= 0 && \text{MLEs } \hat{\sigma}^2 \text{ and } \hat{\mu} \text{ satisfy this eqn} \\
\hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{X})^2
\end{aligned}$$

Notice that the MLE of the variance is *not* the sample variance. In fact, the MLE $\hat{\sigma}^2$ is *biased*.

**MLE of Variance is Biased**

To verify the bias of the variance requires some algebra.

$$\begin{aligned}
E[\hat{\sigma}^2] &= E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} E\left[X_i^2 - \frac{2}{n}X_i\sum_{j=1}^{n}X_j + \frac{1}{n^2}\left(\sum_{j=1}^{n}X_j\right)^2\right] \\
&= \frac{1}{n}\left[\sum_{i=1}^{n}E[X_i^2] - \frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}E[X_iX_j] + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}E[X_iX_j]\right] \\
&= \frac{1}{n}\left[n(\sigma^2+\mu^2) - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}(\text{Cov}(X_i,X_j) + E[X_i]E[X_j])\right] \\
&= \frac{1}{n}\left[n(\sigma^2+\mu^2) - \sigma^2 - n\mu^2\right] \\
&= \frac{n-1}{n}\sigma^2 \neq \sigma^2
\end{aligned}$$

### 1.3.2 Confidence Intervals

#### Point Estimates

We now have a very generic method for finding population parameter estimates $\hat{\theta}$. We have already seen that it is not perfect. MLEs can be biased. There is no one universal best estimator for any parameter, which is, in part, why statisticians are still employed. Maximum likelihood estimators, despite their faults, are intuitive, conceptually easy to find (if not always practically easy), and universal (apply to any problem). For this reason, they are one of the most important estimators to learn.

However, ML estimators and all other estimators like them are only *point estimates* of the population parameter $\theta$. They are one number $\hat{\theta}$ meant to be close to the truth $\theta$. But, they are also random variables. Another dataset would provide another estimate, and as random variables, they come with a variance. We need additional summaries to interpret the value of any point estimate that we obtain.

#### Confidence Interval

Definition: *confidence interval*

> Suppose $\hat{\theta}$ estimates $\theta$. If
>
> $$P(|\hat{\theta} - \theta| < e) > \alpha,$$
>
> then $(\hat{\theta} - e, \hat{\theta} + e)$ is an $\alpha{\cdot}100\%$ confidence interval (CI) of $\theta$. Useful CI are obtained for $\alpha = 0.9, 0.95, 0.99, 0.999$, *etc.*.

Test yourself with some questions.

- If estimator $\hat{\theta}_A$ is more efficient than estimator $\hat{\theta}_B$, which has smaller CI?

  $$\hat{\theta}_A$$

- If estimator $\hat{\theta}_A$ is a consistent estimator and I take a sample of size $n = 100$ and another with size $n = 1000$, which sample gives the smaller confidence interval?

  The sample with size $n = 1000$

- If an estimator $\hat{\theta}_A$ is biased and estimator $\hat{\theta}_B$ is not, which has smaller CI?

  It depends on the balance between bias and efficiency.

**Interpreting CI**

Interpreting CI can be tricky. You should draw a picture to follow along with this discussion. Try explaining it to someone else to test your comprehension.

Let us consider the sample mean $\bar{X}$ as an esimator of the population mean $\mu$. $\bar{X}$ has a sampling distribution (approximately normal) with mean $\mu$. Because we know this sampling distribution, we can find $e$ such that $(\mu - e, \mu + e)$ is $\alpha$·100% likely to contain the sample mean $\bar{X}$ of our next data collection experiment. Now, suppose it turns out that $\bar{X} < \mu$ and $\mu - \bar{X} = a < e$ (notice, there was an $\alpha$·100% chance that $a < e$). Then, for this particular dataset $\mu \in (\bar{X} - e, \bar{X} + e)$. Occassionally, with probability $(1 - \alpha)$, we will collect data such that $\bar{X} + e < \mu$ or $\bar{X} - e > \mu$. For this data, $\mu \notin (\bar{X} - e, \bar{X} + e)$. When this happens, we will not know it, but if our theory is correct, the chance of such an occurrence is small $(1 - \alpha)$.

We should *never* conclude that $\mu$ is in $(\bar{X} - e, \bar{X} + e)$ with probability $\alpha$. Population parameter $\mu$ is a constant, albeit unknown. It is not a random variable, so it does not make sense to associate probability statements with it. The probability statement we can make is with regard to the random variable $\bar{X}$ of an as-yet unseen data set. However, once the data set has been observed and $\bar{X}$ computed, then $\mu \in (\bar{X} - e, \bar{X} + e)$ or it isn't. End of story.

**Large Sample Approximate CI for Population Mean $\mu$**

For large samples (approximately $n \geq 30$), we can construct approximate confidence intervals for the population mean $\mu$ by using the CLT. If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F(x; \mu, \sigma^2)$, then $\bar{X} \sim N(\mu, \sigma^2)$ approximately by CLT. Further,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Let $\Phi(x)$ be the cdf of a standard normal. Then, we can construct approximate CI as follows We seek $e$ such that

$$\begin{aligned}
P(|\bar{X} - \mu| < e) &\geq \alpha \\
P\left(\frac{|\bar{X} - \mu|}{\sigma/\sqrt{n}} < \frac{e}{\sigma/\sqrt{n}}\right) &\geq \alpha \\
P\left(|Z| < \frac{e}{\sigma/\sqrt{n}}\right) &\geq \alpha \\
P\left(-\frac{e}{\sigma/\sqrt{n}} < Z < \frac{e}{\sigma/\sqrt{n}}\right) &\geq \alpha \\
\Phi\left(\frac{e}{\sigma/\sqrt{n}}\right) - \Phi\left(-\frac{e}{\sigma/\sqrt{n}}\right) &\geq \alpha \\
2\Phi\left(\frac{e}{\sigma/\sqrt{n}}\right) - 1 &\geq \alpha \\
\Phi\left(\frac{e}{\sigma/\sqrt{n}}\right) &\geq \frac{\alpha + 1}{2} \\
\frac{e}{\sigma/\sqrt{n}} &\geq \phi_{\frac{1+\alpha}{2}} \\
e &\geq \phi_{\frac{1+\alpha}{2}} \frac{\sigma}{\sqrt{n}}
\end{aligned}$$

In this derivation, we have used basic probability, the CLT, and definition of quantile. To obtain $\alpha$·100% CI, then we need to choose $e$ at least as big as $\phi_{\frac{1+\alpha}{2}} \frac{\sigma}{\sqrt{n}}$. The $\alpha$·100% CI are at least as wide as

$$\left(\bar{X} - \phi_{\frac{1+\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + \phi_{\frac{1+\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right)$$

**Example: Salaries**

Example:

The goal is to estimate the 95% CI for mean salary of an ISU employee when $\sigma = 3000$ is known. To accomplish this goal, 100 ISU employees are randomly sampled and the sample mean is $\bar{X} = 21543$.

The desired CI is

$$\left(21543 - 1.96 \times \frac{3000}{\sqrt{100}}, 21543 + 1.96 \times \frac{3000}{\sqrt{100}}\right)$$

where we find the value 1.96 from `R` as `qnorm((0.95+1)/2)`.

To reiterate an old point, notice that if we repeated the experiment, the CI would change. If we were to randomly sample 100 ISU employees again and again, then we expect 95% of the constructed CI to include the true mean $\mu$ salary. This is only approximate because it is unclear exactly when the CLT kicks in, however with $n = 100 > 30$, we are quite confident that $\bar{X}$ has a near normal sampling distribution.

**Example: simulator**

Example:

You have a complicated model of a queueing system. There are no formulae or theory, but you've written code that simulates it. Because the model is probabilistic, different simulation runs (with different seeds) produce different results. You want to know the mean queue length after 1000 hours (the long-run queue length). You run the simulator 50 times and obtain data with sample mean $\bar{X} = 21.5$ and sample standard deviation $s = 15$. What are 90% CI for the long-run queue length of this model?

The desired CI are

$$21.5 \pm 1.64 \frac{15}{\sqrt{50}}$$

where we obtained the quantile as `qnorm((1+0.90)/2))`.

Note, to solve this example, we had to substitute in an estimate of the population standard deviation $\sigma$. Because the estimator $s$ is a random variable, and not perfect, it means our coverage probability may be slightly off. If coverage probability is off, it means the CI from a random experiment may not achieve $\alpha \cdot 100\%$ chance of containing the true population mean $\mu$.

**Example: population proportion**

Example:

Let $p$ be a proportion of a population satisfying some property (e.g. has disease). Sample $n$ individuals and record how many $X$ have the property.

Notice $X = Y_1 + \cdots + Y_n$ is a sum of iid Bernoulli random variables, so by CLT,

$$\frac{X}{n} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

Immediately, we notice that the estimator $\hat{p} = \frac{X}{n}$ is unbiased for population proportion $p$. Furthermore, we can compute CI as

$$\hat{p} \pm \phi_{\frac{1+\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

**Population Proportion CI**

In this case, we will never know the population proportion $p$. (If we knew it, then why are we wasting our time estimating it with $\hat{p}$ and CI?) There are multiple ways we could replace the unknown $p$ in the CI formula.

- **Conservative Method.** $p(1-p)$ is maximized when $p = 0.5$, so assume maximum variance to construct CI
$$\hat{p} \pm \phi_{\frac{1+\alpha}{2}} \frac{1}{2\sqrt{n}}$$
- **Substitution Method.** Replace $p$ with estimate $\hat{p}$.
$$\hat{p} \pm \phi_{\frac{1+\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

There will be very little difference between these methods with $n$ is large or $p \approx 0.5$.

**Example: queueing system revisited**

Example:

> Revisit the queueing system and find 95% CI for the probability that a server is available after a long time. Simulate 100 times and determine if there is a server available at time $t = 1000$ hours. The data finds 65 of 100 have a free server.
>
> $\hat{p} = 0.65$ and the CI are, by conservative method,
> $$0.65 \pm \frac{1.96}{20} = 0.65 \pm 0.098$$
> or, by substitution method,
> $$0.65 \pm 1.96\sqrt{\frac{0.65 \cdot 0.35}{100}} = 0.65 \pm 0.093$$

### 1.3.3 Generalizing the CLT

**Generalizing CLT**

We will now spend some time generalizing the CLT to more complex situations so that we can continue our discussion of CI.

**Theorem 3.** *Suppose $X_1, \ldots, X_n$ are independent with $X_i \sim N(\mu_i, \sigma_i^2)$. Form $S_n = \sum_{i=1}^n X_i$, then $S_n \sim N(\mu, \sigma^2)$ where*

$$\begin{aligned} \mu &= \sum_{i=1}^n \mu_i \\ \sigma^2 &= \sum_{i=1}^n \sigma_i^2 \end{aligned}$$

**Corollary 4.** *Suppose $X_1, \ldots, X_n$ are as in the theorem and $a_1, \ldots, a_n$ are constants. Then $Y_n = \sum_{i=1}^n a_n X_i$ is normally distributed with*

$$\begin{aligned} \mu &= \sum_{i=1}^n a_i \mu_i \\ \sigma^2 &= \sum_{i=1}^n a_i^2 \sigma_i^2 \end{aligned}$$

**Example: linear combinations of normals**

Example:

Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, then

$$\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$$

Notice, this implies that the CLT is *exactly* true for data $X_1, \ldots, X_n$ that is iid Normal. The more normal the data sampling distribution, the more exact the asymptotic distribution of the CLT is.

Example:

Suppose $X_1 \sim N(\mu_1, \sigma_1^2)$ and independently $X_2 \sim N(\mu_2, \sigma_2^2)$, then

$$X_1 - X_2 \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$$

### 1.3.4 CI for Differences

**CI for Differences**

With this background in hand, we are ready to construct CI for differences of population means and proportions.

Example:

Suppose you have two samples.

$$\begin{aligned} X_{11}, \ldots, X_{1n_1} &\overset{\text{iid}}{\sim} N(\mu_1, \sigma_1^2) \\ X_{21}, \ldots, X_{2n_2} &\overset{\text{iid}}{\sim} N(\mu_2, \sigma_2^2) \end{aligned}$$

Use these data to construct CI for $\mu_1 - \mu_2$.

A natural estimator is $\bar{X}_1 - \bar{X}_2$. But by the extensions of the CLT, we know, asymptotically,

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

Thus, the CI are

$$\bar{X}_1 - \bar{X}_2 \pm \phi_{\frac{1+\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

**Example: CS vs. Other Majors**

Example:

Suppose $X_1, \ldots, X_{11}$ are the review #1 exam scores for the CS majors and $Y_1, \ldots, Y_{14}$ are the corresponding scores for non-CS majors. What is the 95% CI for the difference in mean exam scores of all CS majors vs. all non-CS majors.

We assume that the members of the class are random samples from the CS major pool and the non-CS major pool at ISU. We compute $\bar{X} - \bar{Y} = 6.55$, which suggests that CS majors score better, onaverage, on the probability review exam. Of course, we need to consider variability in the data. We summarize it via CI. $s_X = 176.69$ and $s_Y = 267.17$, so $\text{Var}(\bar{X} - \bar{Y}) = 35.15$. Then, the CI is

$$(-5.07, 18.17)$$

and we cannot conclude with high confidence that CS majors generally score better on this probability review exam.

**Example: proportions**

Example:

Suppose there are two samples, each yielding a different estimate $\hat{p}_1$ and $\hat{p}_2$ of a proportion. We seek CI for the population difference $p_1 - p_2$.

By the CLT, we know $\hat{p}_i \sim N\left(p_i, \frac{p_i(i-p_i)}{n_i}\right), i = i, 2$, and therefore

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

yielding CI

$$\hat{p}_1 - \hat{p}_2 \pm \phi_{\frac{1+\alpha}{2}} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Again, we have two choices for filling in the unknown $p_1$ and $p_2$.

- **Conservative method.** Set $p_1 = p_2 = \frac{1}{2}$ to yield

$$\hat{p}_1 - \hat{p}_2 \pm \frac{\phi_{\frac{1+\alpha}{2}}}{2} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- **Substitution method.** Set $p_i = \hat{p}_i, i = 1, 2$ to yield

$$\hat{p}_1 - \hat{p}_2 \pm \phi_{\frac{1+\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

### 1.3.5 One-Sided Confidence Intervals

**One-Sided Confidence Intervals**

Definition: *one-sided confidence interval*

A one-sided $\alpha$·100% CI is $(\hat{\theta} - e, \infty)$ such that

$$P(\hat{\theta} > \theta - e) > \alpha$$

or $(-\infty, e)$ such that

$$P(\hat{\theta} < \theta + e) > \alpha$$

One-sided CI are constructed in just the same way as two-sided CI except you use $\phi_\alpha$. Namely

$$(\hat{\theta} - \phi_\alpha \sigma_{\hat{\theta}}, \infty)$$

or

$$(-\infty, \hat{\theta} + \phi_\alpha \sigma_{\hat{\theta}})$$

where $\sigma_{\hat{\theta}}$ is your estimate of the sampling standard deviation of estimator $\hat{\theta}$.

**Example: CS vs. Other Majors**

Example:

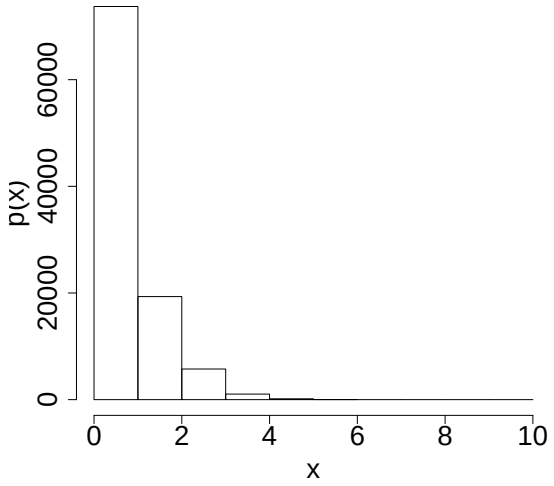



Figure 3: Binomial$(10, 1/10)$ Distribution

Suppose you *know* that CS majors do better than other majors on probability review exams, so you *know* $\mu_x - \mu_y > 0$. Put a lower bound on $\mu_x - \mu_y$ as an indication of *how much better* they perform.

We estimate CI $(\bar{X} - \bar{Y} - e, \infty)$ with lower bound

$$6.55 - 1.64\sqrt{\frac{176.69}{11} + \frac{267.17}{14}} = -3.20.$$

Perhaps we should rein in our confidence. The data provide no more detailed information than we already knew: $\mu_x - \mu_y > 0$.

## 1.4 Hypothesis Testing

**Hypothesis Testing**

We have covered the use of statistics to estimate population parameters, including population mean, population proportion, and differences in both of these. We now consider the second objective of statistics: to test hypotheses about the population. We motivate this section with an example.

Example:

A divining rod is a metal or wooden rod used to find ground water, oil, other riches beneath the ground without digging. A dowser is an individual who operates a divining rod.

Suppose a dowser is asked to identify which of 10 pipes, buried beneath the ground, is transporting water. The experiment is set up such that one and only one of the pipes is transporting water at any one time. He attempts 10 times and finds the pipe with water 3 times. Does he have special skills?

The objective is to test whether the dowser has skills. We start with a hypothesis about the dowser, a simple one so that we can formulate a sampling distribution under this hypothesis. We assume the dowser has no skills, so he is randomly selecting the pipe. What is the sampling distribution for $X = 3$, the number of times he correctly guessed the pipe?

$$X \sim \text{Binomial}(10, 1/10)$$

Why did we choose the hypothesis we chose? There really was no other choice. If we had started with a more complicated hypothesis, i.e. the dowser has skills, we would have quickly run into problems. What skills, exactly, does the dowser have? If you say he always gets the right pipe, then you immediately disprove the hypothesis, since $X \neq 10$. Then what? He gets the pipe right $30\%$ of the time, but what is the basis for your choice $30\%$? For this reason, we choose simpler hypotheses until we are forced, by data, to choose more complicated hypotheses. Also, in the scientific method, we can never prove a hypothesis correct, so if we start with the presumption that he has skills, we could never prove it! (Although, we certainly could collect lots of corroborating evidence, slowly turning our hypothesis into a theory, and eventually into a law.)

Fig. 3 shows the probability mass function for our hypothesized distribution of $X$. We use this distribution to assess the validity of our hypothesis. We do so by determining whether the observed data $X = 3$ is unusual. What is the probability, if the hypothesis is correct, that we observe data as or more unusual than $X = 3$? The probability of $P(X = 3) = 0.05739563$, according to `dbinom(3, size=10, prob=1/10)`. Clearly, events $\{X = 4\}, \{X = 5\}, \ldots$ are all more unusual (less likely). Thus, the probability of observing data as or more unusual is

$$P(X \geq 3) = 1 - P(X \leq 2) = 0.07019083$$

computed as `1-pbinom(2, size=10, prob=1/10)`.

This result shows that our dowser has done something fairly unusual. Typically, by convention, however, we look for events that are less than 5% probable before declaring our hypothesis invalid. (More about this kind of decision later.)

**Hypothesis Testing Procedure**

1. Formulate a null hypothesis about a population parameter of the form

   $$H_0 : f(\theta) = A$$

   for some constant $A$, population parameter $\theta$ and function $f(\cdot)$.

2. Identify the alternate hypothesis that you would consider if $H_0$ is found to be unlikely. Typically, the hypothesis is one of

   $$H_A : f(\theta) > A$$
   $$H_A : f(\theta) < A$$
   $$H_A : f(\theta) \neq A$$

   and most commonly it is the last one.

3. Gather data $X$ and compute a test statistic $t(X) = t$. Derive the (approximate) sampling distribution for r.v. $t(X)$ when $H_0$ is true.

4. Compute the probability, assuming $H_0$, of getting $t(X)$ as extreme or more extreme than the observed $t$.

5. Reject $H_0$ if the probability computed above is too low.

**Example: sloppy coders**

A student job queue is configured to optimally handle jobs that run for 2 minutes. One faculty suspects that students have gotten sloppier in how they code, so that run lengths now average more than 2 minutes. The faculty decide that if this is true, they will change the queue parameters *and* force students to take a 90-minute coding class to learn the art of efficient programming. The faculty ask the computer support person to randomly samples 50 jobs. She does and reports $\bar{X} = 2.09$ and $S = 0.31$ minutes. Is there evidence that mean run length has increased?

1. We establish the null hypothesis as
$$H_0 : \mu = 2 \text{ minutes}$$
2. For this case, we are only interested in the alternative
$$H_A : \mu > 2$$
because we will do nothing if $\mu \leq 2$.
3. We know that $\bar{X} \sim N(\mu, \sigma^2/n)$, where $\mu$ is the mean job length and $\sigma^2$ is the job length variance. Under $H_0$, $\mu = 2$, but variance is still unknown. Under $H_A$, we expect $\bar{X}$ to be large. Thus, if we use $t(X) = \bar{X}$, it is high values of $\bar{X}$, larger or equal to the observed $2.09$, that are unusual.
4. We compute
$$P(\bar{X} > 2.09) = 0.02004159$$
by `1-pnorm(2.09, mean=2, sd=0.31/sqrt(50))`.
5. We reject $H_0$ and conclude that job lengths have indeed increased since the queue was originally established.

**Quick List of Hypothesis Tests**

A quick list of hypothesis test which you can perform, based on the theory you know so far, are listed below.

| Hypothesis | Null Distribution |
|---|---|
| $H_0 : \mu = A$ | $\bar{X} \sim N(A, \sigma^2/n)$ |
| $H_0 : p = A$ | $\hat{p} \sim N(A, A(1-A)/n)$ |
| $H_0 : \mu_1 - \mu_2 = A$ | $\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)$ |
| $H_0 : p_1 - p_2 = 0$ | $\hat{p}_1 - \hat{p}_2 \sim N\left(0, \hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right), \hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1+n_2}$ |

Notice that the variances for the null distributions for tests of proportions, are obtained under the assumption of the null hypothesis. No need to resort to the conservative method or substitution method for variances (as we did when computing CI). Now, the assumption of $H_0$ allows us to plug in $p = A$ for $H_0 : p = A$ and $p = \hat{p}$ for $H_0 : p_1 - p_2 = 0$. For the latter, although we know $p_1 = p_2$, we do not know the value, and must estimate it from the complete data. The numerator in $\hat{p}$ is $n_1\hat{p}_1 + n_2\hat{p}_2$, the total number of successes in both samples.

**Example: long-run server availability**

Returning to the example of complex queues, suppose you have two alternative queues, and you would like to test whether there long-run server availability is the same or different. You simulate queue 1 $n_1 = 1000$ times and find $\hat{p}_1 = \frac{551}{1000}$. The second queue is more complex so you only manage to simulate it $n_2 = 500$ times and $\hat{p}_2 = \frac{303}{500}$.

1. The null hypothesis is $H_0 : p_1 = p_2$.
2. The alternative is any differences $H_A : p_1 \neq p_2$.
3. Under the null hypothesis, $\hat{p}_1 - \hat{p}_2 \sim N\left(0, p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$, which we can obtain by substituting $\hat{p} = \frac{551+303}{1500}$.
4. The probability of observing data this extreme is
$$P\left(\hat{p}_1 - \hat{p}_2 \leq -\frac{55}{1000}\right) = 0.004458117$$
obtained as

```
> p <- (551+303)/1000
> 2*pnorm(-55/1000, mean=0, sd=sqrt(p*(1-p)*(1/500+1/1000)))
```

5. We reject $H_0$ and conclude that there is a difference in the queue long-run availabilities.

**Decision Making**

Hypothesis testing is a process of decision making. We ask whether a hypothesis $H_0$ is true and answer "Yes" or "No." Our decision is dichotomous, but it is based on a continuous probability (probability of as extreme or more extreme than the observed). The probability communicates the fact that we can never be certain about our decision, and in fact there are two types of errors we can make when we draw our conclusion.

| | $H_0$ true | $H_0$ false |
|---|---|---|
| Reject $H_0$ | type I error ($\alpha$) | $1-\beta$ |
| Accept $H_0$ | $1-\alpha$ | power ($\beta$) |

Definition: *type I error*

> If we reject $H_0$ when it is actually true, then we are making an error of type $I$.

We generally want to avoid this kind of error because it means we reject a simpler hypothesis $H_0$ in favor of a more complex hypothesis. We do not want to be forced down the road of complexity without good reason. Thus, hypothesis tests are set up to control type I error, generally setting $\alpha = 0.05$, or less if the situation demands it.

Definition: *type II error*

> If we accept $H_0$ when it is false, then we are committing a type II error.

While we would like to avoid this error too, we are not so worried if it happens because we presume that we will detect the problem later. If we fail to reject $H_0$ when it is not true, we will continue life under the illusion that $H_0$ is correct. Eventually, some contradiction ought to show up to suggest $H_0$ is not true. Of course, a company that wants to claim its product is better than the competitor may care very, very much about this kind of error. We will discuss how to control this error (at some cost $$) in the future.

Definition: *p-value*

> The probability of obtaining a statistic $t(X)$ as extreme or more extreme than the observed $t(X) = t$ is called the $p$-value.

### 1.4.1 More Theory

It is time to dig into some more theory so we can develop our next hypothesis test. First, we review the change-of-variable theorem, and then use it to prove our next important distribution, the distribution of $Z^2$, where $Z \sim N(0,1)$.

**Change-of-Variable Theorem**

**Theorem:** change-of-variable

Let $X$ have pdf $f_X(x)$ and $Y = g(X)$ where $g(\cdot)$ is monotone. Let $\Sigma_X = \{x : f_X(x) > 0\}$ and $\Sigma_Y = \{y : y = g(x) \text{ for some } x \in \Sigma_X\}$. Suppose $f_X(x)$ is continuous on $\Sigma_X$ and $g^{-1}(\{y\})$ has continuous derivative on $\Sigma_Y$. Then,

$$f_Y(y) = \begin{cases} f_X\left[g^{-1}(y)\right] \left|\frac{dg^{-1}(y)}{dy}\right| & y \in \Sigma_Y \\ 0 & \text{otherwise} \end{cases}$$

**Example: change-of-variable**

Example:

Suppose $Z \sim N(0,1)$. Let $Y = g(Z) = \sigma Z + \mu$. Find the distribution of $Y$.

$$Z = g^{-1}(Y) = \frac{Y - \mu}{\sigma}$$

and

$$\frac{dg^{-1}(y)}{dy} = \frac{1}{\sigma}$$

Therefore,

$$f_Y(y) = f_Z\left(\frac{y-\mu}{\sigma}\right)\left|\frac{1}{\sigma}\right| = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]$$

**Chi-Square Distribution**

A special case of the gamma distribution is the chi-square distribution.

**Definition:** $X \sim \chi^2_p$ is said to have a chi-square distribution with $p$ degrees of freedom if it has the Gamma pdf with $\alpha = p/2$ and $\beta = 2$. In other words, it has pdf

$$f_X(x) = \frac{1}{\Gamma(p/2)2^{p/2}}x^{p/2-1}e^{-x/2}.$$

**Extended Change-of-Variable**

**Theorem:** change-of-variable for piecewise monotonic functions

Suppose $X \sim f_X(x)$ and $Y = g(X)$. Suppose $A_0, A_1, \ldots, Z_k$ cover $\Sigma_X$ such that $P(X \in A_0) = 0$ and $f_X(x)$ is continuous on each $A_i$. Further, suppose

1. $g(x) = g_i(x)$ for $x \in A_i$,
2. $g_i(x)$ is monotone on $A_i$,
3. $\Sigma_Y = \{y : y = g_i(x) \text{ for some} x \in A_i\}$ is the same for all $A_i$, and
4. $g_i^{-1}(y)$ has continuous derivative on $\Sigma_Y$ for all $i$.

Then,

$$f_Y(y) = \begin{cases} \sum_{i-1}^{k} f_X\left[g_i^{-1}(y)\right]\left|\frac{d}{dy}g_i^{-1}(y)\right| & y \in \Sigma_Y \\ 0 & \text{otherwise} \end{cases}$$

**Example: Chi-Square**

**Lemma:**

1. $Z \sim N(0,1)$, then $Z^2 \sim \chi^2_1$
2. Reproductive property for $\chi^2$: $X_1, \ldots, X_n$ independent and $X_i \sim \chi^2_{p_i}$, then $X_1 + \cdots + X_n \sim X^2_{p_1+\cdots p_n}$

**Proof:** of 1

Note that $y = g(z) = z^2$ is monotone on $(-\infty, 0)$ and $(0, \infty)$, so let $\Sigma_Y = (0,\infty), A_0 = \{0\}, A_1 = (-\infty, 0), A_2 = (0,\infty)$. Then, $g_1^{-1}(y) = -\sqrt{y}$ on $A_1$ and $g_2^{-1}(y) = \sqrt{y}$ on $A_2$. Therefore,

$$f_Y(y) = \frac{1}{\sqrt{2\pi}}e^{(-\sqrt{y})^2/2}\left|\frac{-1}{2\sqrt{y}}\right| + \frac{1}{\sqrt{2\pi}}e^{-(\sqrt{y})^2/2}\left|\frac{1}{2\sqrt{y}}\right| = \frac{1}{\sqrt{2\pi y}}e^{-y/2}$$

which is $\chi^2_1$.

**Example: more on location/scale transformations**

We have already seen how standard normal $Z$ can be transformed to $Y = Z\sigma + \mu$ to obtain a $N(\mu, \sigma^2)$ distribution. These kinds of transformations are so common, we focus on some general results. Let $f(\cdot)$ be a pdf, $\mu \in \Re$, $\sigma > 0$, then

$$\begin{array}{lll} X = \sigma Z + \mu \text{ and } Z \sim f(z) & \implies & X \sim \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) \\ X \sim \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) & \implies & \exists \text{ r.v. } Z \sim f(z) \text{ and } X = \sigma Z + \mu \end{array}$$

This result allows us to quickly derive the pdf after location/scale transformations (i.e. $X = \sigma Z + \mu$, where $\sigma$ is adjustment in scale, $\mu$ is adjustment in location). It also shows how a certain form of the pdf allows us to infer the existence of another random variable that is a mere location/scale transformation of our existing random variable.

### 1.4.2 Goodness-of-Fit Test

**Goodness-of-Fit Test**

So far, we have used hypothesis testing to make decisions about single parameters, population means, population proportions, population mean differences, etc. Sometimes, however, we will want to test whether a distributional assumption is valid. For example, if we have assumed our data is normal (so that the CLT kicks in even for $n < 30$), then perhaps we need to verify this assumption.

The null hypothesis in this case is

$$H_0 : X_1, \ldots, X_n \overset{\text{iid}}{\sim} F(x; \theta)$$

with alternative

$$H_0 : X_1, \ldots, X_n \overset{\text{iid}}{\sim} G(x; \theta)$$

where $G$ is not $F$.

We will elaborate the discrete, finite case here, where the CDF $F(x; \theta)$ is characterized by a pmf $p(i) = P(X = i)$. Our null hypothesis becomes

$$H_0 : p(i) = p_i \text{ for all } i = 1, \ldots, k$$

and the alternative is

$$H_0 : \exists i \text{ such that } p(i) \neq p_i$$

Example:

> Suppose $N$ reads of length $L = 34$ are randomly taken of a genome. Suppose further that errors are introduced into the reads by independently mutating each nucleotide with constant probability $p_e$. A student designs a method for simulating this process. He proposes to simulate exactly $NLp_e$ errors by randomly distributing them throughout the $NL$ sequenced nucleotides. An advisor suggests that he is not properly simulating the error process because the total number of errors will not be exactly $NLp_e$. As an alternative, the advisor suggests the student generates $X \sim$ Binomial$(NL, p_e)$, and then distributes the $X$ random mutations uniformly through the $NL$ positions.
>
> Both procedures are implemented on a small dataset and the number of reads with 0, 1, 2, up to 34 errors are recorded.

| Method | 0 | 1 | 2 | 3 | $\cdots$ | 34 |
|---|---|---|---|---|---|---|
| His Simulation | 856 | 133 | 11 | 0 | $\cdots$ | 0 |
| Advisor's Simulation | 842 | 147 | 11 | 0 | $\cdots$ | 0 |
| Expected | 843.31 | 144.08 | 11.95 | 0.64 | $\cdots$ | 0 |

> The idea is to test whether either of these simulations has data that contradicts the assumption of independent errors with constant probability $p_e$.

1. The probability of 0 errors under the null model is $p_0 = \binom{L}{0} p_e^0 (1-p_e)^L$. The probability of 1 error is $p_1 = \binom{L}{1} p_e (1-p_e)^{L-1}$. And so forth, up to $L$ errors.
2. The alternative is that any one of these probabilities is not satisfied.
3. We need a statistic that is sensitive to violation of the assumptions. Let $X_0, X_1, \ldots, X_{34}$ be the observed counts in each category. We propose

$$t(X) = \sum_{i=0}^{34} \frac{(X_i - E_i)^2}{E_i}$$

where $E_i = E[X_i \mid H_0]$. By the CLT, $X_i \sim N(np_i, np_i(1-p_i))$ under the null hypothesis, and

$$\frac{X_i - np_i}{\sqrt{np_i(1-p_i)}} \sim N(0,1)$$

Since $E_i = np_i$, the terms $\frac{(X_i - E_i)^2}{E_i}$ are very nearly (though not quite) squares of standard normal random variables. By the reproductive property of chi-squared random variables, we conclude $t(X) \sim \chi^2_{34-1}$. The degrees of freedom we determine by computing the expectation (remember the expectation of chi-squared r.v. is the d.f.)

$$\begin{aligned}
E[t(X)] &= \sum_{i=0}^{34} E\left[\frac{(X_i - E_i)^2}{E_i}\right] \\
&= \sum_{i=0}^{34} \frac{1}{np_i} E[(X_i - np_i)^2] \\
&= \sum_{i=0}^{34} \frac{1}{np_i} \left\{ \text{Var}((X_i - np_i)^2) + E[X_i - np_i]^2 \right\} \\
&= \sum_{i=0}^{34} \frac{np_i(1-p_i)}{np_i} \\
&= \sum_{i=0}^{34} (1-p_i) = 35 - 1
\end{aligned}$$

One problem with our analysis is that we have relied on the CLT to conclude each of our terms are approximately $N(0,1)$. The CLT is appropriate for the Binomial distribution so long as $p$ is neither too large nor too small, but evidently $p_i \approx 0$ for $i > 2$. Thus, we may not rely on the CLT for our asymptotics *unless* each category is fairly probable. The general rule of thumb is that $E_i < 5$ for no more than 10% of the categories $i$. Since the rule of thumb is not satisfied for our data, we get around the problem by combining several categories together, in particular merging all categories for $i \geq 2$. The new expected value for this merged category is $E_2 = 12.61$.

The probability of data as or more extreme than observed is

$$P(t(X) > t \mid H_0)$$

where $t$ is the observed statistic.

In our case, $t_H(X) = 1.25$ for his method and $t_H(X) = 0.267$ for the advisor's method. The associated probabilities are $0.5352614$ and $0.8750275$ obtained with `1-pchisq(1.25, df=2)` and `1-pchisq(0.267, df=2)`.

4. We can reject the null hypothesis for neither method.

**Note.** There was much discussion in class about the whether to perform a one-sided or two-sided test. Generally, a chi-square test is performed as a one-sided test. You expect the

statistic to get large as the alternative distribution $G(x;\theta)$ deviates from $F(x;\theta)$. Thus extreme values are in the right tail. Sometimes, however, you may be concerned that the data are unexpectedly close to the hypothetical distribution. This could happen if your random number generator is not working, or if you have somehow erased normal randomness that is supposed to be there. In this case, you might also be concerned about the area under the left tail, for exceptionally small statistics. Note, the statistic is always positive, so the left tail is right up against $0$.

**Continuous Random Variable**

We discussed how one would perform goodness-of-fit tests for continuous random variables. Bin the observations into bins, such that the probability of each bin is not too small. Then continue as discussed above.