2 Classical Statistical Model. Statistics

2.1 Probability Versus Statistical Inference

Complementary processes:

- Probability Theory: We start with a completely specified model, which we assume as "correct" and we compute probabilities of certain events;
- Statistical inference: We observe the realization of certain events, and using that information we try to infer the probabilistic model that governs the corresponding random experiment.

Example 2.1 *Consider a very large group of people. Assume that the proportion of smokers is* θ.

- *With known* θ, *we can use probability theory to compute the probability that when selecting* n *people at random from this group, we find exactly* x *smokers;*
- *In practice, it is almost always the case that* θ *is unknown. We select 10 people at random from this population and observe three smokers. From this information, we want to make inferences about* θ. *This is a problem of Statistical Inference.*

■

- Statistical data result from experiments conducted on a subset of a population — the sample — and we try to extend the conclusions obtained to the whole population
- Typical diagram:

$$\begin{array}{ccc} & \boxed{\text{Population}} & \\ \text{Sampling} & \downarrow \quad \uparrow & \text{Inductive inference} \\ & \boxed{\text{Sample}} & \end{array}$$

2.2 Model Specification. Random Sample

Formalizing the process of statistical inference

- Characteristic of interest is modeled as a random variable X with cdf F — the *statistical model*
- That model must be specified:
  - parametric models — F i known up to a finite dimensional parameter (k-dimensional, say). For instance, we can model X as normal with mean μ and variance σ^2, both unknown
  - nonparametric models — F is specified in a nonparametric fashion, e.g., F is an element of the set of all continuous and symmetric distributions.

The case that interests us the the parametric one. Parametric statistical model:

$$\mathcal{F} = \{F(\cdot \mid \theta) : \theta \in \Theta\}$$

The set Θ is known as the parametric space.

Example 2.2 *If we are interested in the daily return of a financial asset we can propose a normal or a gamma:* $\mathcal{F} = \{\mathrm{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma > 0\}$*, or* $\mathcal{F} = \{\mathrm{G}(\alpha, \lambda) : \alpha, \lambda > 0\}$

If interested in studying the number of claims per year in an insurance policy, we can resort to the Poisson distribution, $\mathcal{F} = \{\mathrm{Po}(\lambda) : \lambda > 0\}$*.* ■

The specification stage is very important and results from many factors, namely

- knowledge of the problem at hand
- knowledge of previous studies
- knowledge of probability theory

The consequences of model misspecification are always negative, but typically its impact is smaller for larger samples

Sampling

- One can imagine many different ways of collecting data
- Random sampling: the observed data are one of many possible data sets we could have obtained in the same circumstances. The set of n observations, $(x_1, \ldots, x_n)$, which we have observed is a realization of an n-dimensional random variable $(X_1, \ldots, X_n)$:

$$\begin{array}{ll} (X_1, \ldots, X_n) & \text{Random sample} \\ (x_1, \ldots, x_n) & \text{Observed sample} \end{array}$$

- Sample space: subset of $\mathbb{R}^n$ that contains the set of possible values for $x_1, \ldots, x_n$. We denote it by $\mathcal{X}$.
- In this course, we limit ourselves almost exclusively to a particular sampling process:

Definition 2.1 IID random sampling: *When the n random variables that compose the random sample are*

1. *mutually independent*
2. *identically distributed, with the same distribution as X*

we say that $(X_1, \ldots, X_n)$ constitutes and iid random sample of size n obtained from the population X. In symbols, $X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} X$. ■

If $\mathcal{F} = \{f(\cdot \mid \theta) : \theta \in \Theta\}$ e $X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} X$, then

$$\begin{aligned} F_{X_1,\ldots,X_n}(x_1, \ldots, x_n \mid \theta) &= \prod_{i=1}^{n} F_{X_i}(x_i \mid \theta) \quad \text{by independence} \\ &= \prod_{i=1}^{n} F(x_i \mid \theta) \quad \text{since } X_i \sim X \end{aligned}$$

and similarly for the probability (density) function:

$$f(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

Example 2.3 *If $X_1, \ldots, X_n$ is an iid random sample from a $Po(\lambda)$ population, then*

$$P(X_1 = x_1, \ldots, X_n = x_n \mid \lambda) = f(x_1, \ldots, x_n \mid \lambda) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!},\ x_i \in \mathbb{N}_0$$

If $X_1, \ldots, X_n$ is an iid random sample from a $N(\mu, 1)$ population, then

$$f(x_1, \ldots, x_n \mid \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x_i - \mu)^2\right] = (2\pi)^{-n/2}\ \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right],$$

$$x_i \in \mathbb{R}$$

■

2.3 Statistics

Definition 2.2 Statistic *A statistic is any function of* $(X_1, \ldots, X_n)$ *that does not depend on unknown parameters.* ■

Example 2.4 *In the context of a* $N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$ *and* $\sigma > 0$ *unknown, examples of unidimensional statistics include*

$$T = \sum_{i=1}^{n} X_i, \quad \bar{X} = \frac{1}{n} T, \quad S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

and examples of bidimensional statistics include

$$\left(T, \sum_{i=1}^{n} X_i^2 \right), \quad (\bar{X}, S^2) \ .$$

The following quantities are not statistics

$$\sum_{i=1}^{n} (X_i - \mu)^2, \quad \frac{1}{\sigma^2} \sum_{i=1}^{n} X_i^2$$

because they depend on unknown parameters. If σ^2 *is known,* $\frac{1}{\sigma^2} \sum_{i=1}^{n} X_i^2$ *is indeed a statistic.* ■

Important examples

- The whole random sample $(X_1, \ldots, X_n)$ is a statistic
- The sample average

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

- The sample variance

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

- The bias-corrected sample variance

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

- the sample maximum, $\max\{X_1, \ldots, X_n\}$
- X_1 or X_n

- Statistics operate a data reduction: clearly observing $(X_1, \ldots, X_n)$ is more informative than observing $\bar{X}$; observing $(\bar{X}, S^2)$ is more informative than observing $\bar{X}$ only

- Statistics are summaries of the information contained in the random sample

- Statistics are random variables. As usual, it is important to distinguish between the random variable and its observed value

population	random sample	observed sample
X	$(X_1, \ldots, X_n)$	$(x_1, \ldots, x_n)$
population mean	sample mean	mean of the sample
$\mu = E[X]$	$\bar{X} = \frac{1}{n}\sum_i X_i$	$\bar{x} = \frac{1}{n}\sum_i x_i$
population variance	sample variance	variance of the sample
$\sigma^2 = \mathrm{Var}(X)$	$S^2 = \frac{1}{n}\sum_i (X_i - \bar{X})^2$	$s^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2$

2.4 Sampling distributions

- The sampling distribution of a statistic corresponds to its probability distribution: as $(X_1, \ldots, X_n)$ varies according to its probability distribution, what is the resulting probabilistic behavior of $T(X_1, \ldots, X_n)$

- In classical inference, it turns out to be very important to known the sampling distribution of statistics because that is necessary to evaluate the performance of statistical methodologies

- Objective: to determine (aspects of) the sampling distribution of a statistic T, knowing (aspects of) the probability distribution of the population X.

Methods to obtain the sampling distribution of a statistic:

- Change of variable: if X is continuous,

$$F_T(t \mid \theta) = P(T \leq t \mid \theta) = \int_{A(t)} \prod_{i=1}^{n} f(x_i \mid \theta) dx_1 \ldots dx_n$$

  where $A(t) = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : T(x_1, \ldots, x_n) \leq t\}$. If X is discrete, replace integrals with sums. Labor-intensive and whenever we can obtain results there are typically more elegant approaches

- Determining the moment generating function of T

- Using well-known properties of the distribution of X (related with the point above)

- Asymptotic approximations to the sampling distribution of certain statistics (CLT and related results)

- Using simulation: very important strategy as the statistical models become more and more complex and the computer power becomes cheaper

Example 2.5 *Let* $T = \sum_{i=1}^{n} X_i$.

- *If* $(X_1, \ldots, X_n)$ *is an iid random sample from a* $Po(\lambda)$ *population, since the sum of independent Poisson is still Poisson, we have* $T \sim Po(n\lambda)$*, hence*

$$f_T(t \mid \lambda) = e^{-n\lambda} \frac{(n\lambda)^t}{t!}, \quad t \in \mathbb{N}_0 \ .$$

- *If* $(X_1, \ldots, X_n)$ *is an iid random sample from a* $N(\mu, \sigma^2)$ *population, then* $T \sim N(n\mu, n\sigma^2)$

- *If* $(X_1, \ldots, X_n)$ *is an iid sample from a* $B(1, \theta)$ *population, then* $T \sim B(n, \theta)$.

$\blacksquare$

2.4.1 Monte Carlo simulation

We have already seen that having a sufficiently large sample drawn from a probability distribution is enough to (approximately) determine many aspects of that distribution.

How do we obtain an iid sample of size N drawn from the sampling distribution of $T = T(X_1, \ldots, X_n)$:

$x_{11}, \ldots, x_{1n}$	$t_1 = T(x_{11}, \ldots, x_{1n})$
$x_{21}, \ldots, x_{2n}$	$t_2 = T(x_{21}, \ldots, x_{2n})$
$\ldots$	$\ldots$
$x_{N1}, \ldots, x_{Nn}$	$t_N = T(x_{N1}, \ldots, x_{Nn})$

- Draw N independent samples of size n from the distribution of X;
- for each of those samples, compute the observed values of the statistic T;
- The N resulting numbers, $(t_1, \ldots, t_N)$, constitute a sample of size N drawn from the sampling distribution of T.

2.4.2 Sample distribution of the sample moments

Definition 2.3 Sample moments *Let* $(X_1, \ldots, X_n)$ *be an iid random sample of size* n *from a population* X*. For* $k \in \mathbb{N}$ *we define the* k*th raw sample moment as*

$$M'_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

and the k*th central sample moment by*

$$M_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k \ .$$

$\blacksquare$

Remark: once again, it is important to distinguish the sample moments, M'_k and M_k, from the population moments, $\mu'_k = E[X^k]$ and $\mu_k = E[(X - E[X])^k]$, and the observed sample moments, $m'_k = \sum_{i=1}^n x_i^k/n$ and $m_k = \sum_{i=1}^n (x_i - \bar{x})^k/n$.

Important special cases: $\bar{X} = M'_1$ and $S^2 = M_2$, the sample mean and the sample variance.

Theorem 2.1 Properties of the sample mean: *If all the moments exist, then*

$$
\begin{aligned}
E[\bar{X}] &= E[X] = \mu \\
Var(\bar{X}) &= \frac{Var(X)}{n} = \frac{\sigma^2}{n} \\
\mu_3(\bar{X}) &= \frac{\mu_3}{n^2} \\
\mu_4(\bar{X}) &= \frac{3\mu_2^2}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3} \ .
\end{aligned}
$$

■

Remarks:

- As long as the moments exist, these results are valid — notice the generality
- The distribution of $\bar{X}$ is centered around μ
- $\lim_{n\to+\infty} \mathrm{Var}(\bar{X}) = 0$

Theorem 2.2 Properties of the sample variance: *If all the moments exist,*

$$E[S^2] = \frac{n-1}{n}\sigma^2$$

$$Var(S^2) = \frac{\mu_4 - \mu_2^2}{n} - 2\frac{\mu_4 - 2\mu_2^2}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3}$$

■

Remarks:

- $E[S^2] < \sigma^2$
- For this reason, we define the **bias-corrected sample variance**

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

Theorem 2.3 Properties of the bias-corrected sample variance: *If all the moments exist,*

$$E[S^{'2}] = \sigma^2$$
$$Var(S^{'2}) = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\mu_2^2\right)$$

■

Theorem 2.4 Properties of central sample moments: *If all the moments exist,*

$$E[M_k] = \mu_k + \mathcal{O}\left(\frac{1}{n}\right)$$
$$Var(M_k) = \frac{c}{n} + \mathcal{O}\left(\frac{1}{n^2}\right)$$

where c *is a constant which involves central population moments of order* $\leq 2k$. ■

Theorem 2.5 Asymptotic distribution of $\bar{X}$: *As long as $Var(X)$ is finite, we have as a direct consequence of the Central Limit Theorem that*

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \sqrt{n}\,\frac{\bar{X}-\mu}{\sigma} \xrightarrow{d} N(0,1)\ .$$

$\blacksquare$

Remarks

- This result is typically used in the form

$$P(\bar{X} \leq x) \approx \Phi\left(\sqrt{n}\,\frac{x-\mu}{\sigma}\right)$$

  that is, $\bar{X} \overset{a}{\sim} \mathrm{N}(\mu, \sigma^2/n)$

- How large is n needed? It depends. In general, unimodality and symmetry have a positive impact on the speed of convergence

- We can derive similar results for other sample moments, but their practical interest is limited

2.4.3 Order statistics

Definition 2.4 Order statistics: *Let* $(X_1, \ldots, X_n)$ *be an iid random sample. The* i*-th order statistic is denoted by* $X_{(i)}$ *e satisfies*

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$$

■

Remarks:

- We also use the terminology "order statistic" to refer to any function of the $X_{(i)}$
- Common order statistics: sample maximum, $X_{(n)}$, sample minimum, $X_{(1)}$, sample median, $M_e = X_{((n+1)/2)}$ if n odd; $M_e = [X_{(n/2)} + X_{(n/2+1)}]/2$ if n even, sample range, $R = X_{(n)} - X_{(1)}$.
- For the most part, we restrict attention to the continuous case
- $(Y_1, \ldots, Y_n) \equiv (X_{(1)}, \ldots, X_{(n)})$ to simplify the notation.

Theorem 2.6 *The order statistics have a joint pdf given by*

$$g(y_1, y_2, \ldots, y_n) = n! \prod_{i=1}^{n} f(y_i), \quad \text{if } y_1 < y_2 < \ldots < y_n \ .$$

If $u < v$, the joint pdf of (Y_u, Y_v) is

$$g_{u,v}(y, z) = \frac{n!}{(u-1)!(v-u-1)!(n-v)!} \times$$
$$[F(y)]^{u-1} \ [F(z) - F(y)]^{v-u-1} \ [1 - F(z)]^{n-v} f(y) f(z) \ , \quad \text{if } y < z.$$

The pdf and cdf of Y_v:

$$g_v(y) = \frac{n!}{(v-1)!(n-v)!}[F(y)]^{v-1}[1 - F(y)]^{n-v} f(y)$$
$$G_v(y) = \sum_{j=v}^{n} \binom{n}{j} [F(y)]^j [1 - F(y)]^{n-j} \ .$$

■

Theorem 2.7 Important special cases—the maximum and the minimum:

$$G_1(y) = 1 - [1 - F(y)]^n; \quad g_1(y) = n\ f(y)\ [1 - F(y)]^{n-1}$$

$$G_n(y) = [F(y)]^n; \quad g_n(y) = n\ f(y)\ [F(y)]^{n-1}$$

$$g_{1,n}(y,z) = n(n-1)[F(z) - F(y)]^{n-2} f(y) f(z)\ , \quad y < z\ .$$

■

Example 2.6 *If* $X \sim Pa(c, \theta)$, *i.e., if with* $\theta > 0$ *and* $c > 0$,

$$f(x) = \theta \ c^{\theta} \ x^{-(\theta+1)} \ , \quad x > c \ ,$$

then $F(x) = 1 - (c/x)^{\theta}$, $x > c$. *Hence, for* $y, z > c$

$$g_1(y) = n\theta \ c^{n\theta} \ y^{-(n\theta+1)} \ \Rightarrow X_{(1)} \sim Pa(c, n\theta)$$

$$g_n(z) = n \ \theta \ c^{\theta} \ z^{-(\theta+1)} \left[1 - \left(\frac{c}{z} \right)^{\theta} \right]^{n-1}$$

Example 2.7 *Recall that if* $X \sim Ex(\lambda)$, *then* $X_{(1)} \sim Ex(n\lambda)$: *since* $F(x) = 1 - \exp(-\lambda x)$, $x > 0$,

$$G_1(x) = 1 - [1 - (1 - \exp(-\lambda x))]^n = 1 - \exp(-\lambda n x) \ , \quad x > 0 \ .$$

$\blacksquare$

Definition 2.5 Sample quantile: *Let $p \in (0,1)$ and $k = np$. Then, the sample quantile of order p is Z_p such that*

$$Z_p = \begin{cases} Y_{\lceil k \rceil} & \textit{if } k \textit{ is not an integer} \\ \frac{Y_k + Y_{k+1}}{2} & \textit{if } k \textit{ is an integer} \end{cases}$$

where $\lceil k \rceil$ is the integer part of k.

■

Theorem 2.8 Asymptotic distribution of the sample quantile of order p: *Let Z_p be the sample quantile of order p of an iid random sample of size n, obtained from a continuous population with density f. Denote by ξ_p the population quantile of order p. If f is continuous and positive at ξ_p, then*

$$\sqrt{n} f(\xi_p) \frac{Z_p - \xi_p}{\sqrt{p(1-p)}} \xrightarrow{d} N(0,1) \; .$$

■

Important special cases: $p = 1/2$, median, and the first and third quartiles, $p = 1/4$ e $p = 3/4$.

Example 2.8 *If* $X \sim N(\mu, \sigma^2)$*, then* $\xi_{1/2} = \mu$*. Then,* $f(\xi_{1/2}) = (2\pi\sigma^2)^{-1/2}$*. Hence,*

$$\sqrt{\frac{2n}{\pi\sigma^2}}(Z_{1/2} - \mu) \xrightarrow{d} N(0,1) \ .$$

$\blacksquare$

2.4.4 A few sampling distributions

Normal population In what follows, let $(X_1, \ldots, X_n)$ be an iid random sample of size n from a $\mathrm{N}(\mu, \sigma^2)$ distribution.

Distribution of the sample mean, $\bar{X}$

- $\bar{X}$ is a linear combination of independent normals, hence it follows a normal distribution
- We know that $E[\bar{X}] = \mu$ and $\mathrm{Var}(\bar{X}) = \sigma^2/n$
- hence

$$\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n) \quad \text{or} \quad \sqrt{n}\,\frac{\bar{X} - \mu}{\sigma} \sim \mathrm{N}(0, 1)$$

Example 2.9 *Suppose that the duration, in minutes, of local telephone calls can be well modeled by a normal distribution with mean 17 and variance 25. Determine the probability that, in a random sample of size n, the average of the durations is between 16 and 18 minutes.*

With $\mu = 17$, $\sigma^2 = 25$, and $\bar{X}$ representing the sample mean, we have that

$$\begin{aligned} P(16 < \bar{X} < 18) &= P\left(\sqrt{n}\,\frac{16-\mu}{\sigma} < \sqrt{n}\,\frac{\bar{X}-\mu}{\sigma} < \sqrt{n}\,\frac{18-\mu}{\sigma}\right) \\ &= P\left(-0.2\sqrt{n} < \sqrt{n}\,\frac{\bar{X}-\mu}{\sigma} < 0.2\sqrt{n}\right) \\ &= 2\Phi(0.2\sqrt{n}) - 1 \ . \end{aligned}$$

How does this probability behave as n increases? What happens as $n \to \infty$? What about $P(14 < \bar{X} < 16)$? ■

Sampling distribution of the bias-corrected sample variance $S^{'2}$

- Clearly, $\sum_{i=1}^{n}(X_i-\mu)^2/\sigma^2 \sim \chi^2(n)$
- Also,

$$\sum_{i=1}^{n}(X_i-\mu)^2=\sum_{i=1}^{n}[(X_i-\bar{X})+(\bar{X}-\mu)]^2=\sum_{i=1}^{n}(X_i-\bar{X})^2+n(\bar{X}-\mu)^2$$

- Hence,

$$\sum_{i=1}^{n}(X_i-\mu)^2/\sigma^2=(n-1)S^{'2}/\sigma^2+n(\bar{X}-\mu)^2/\sigma^2$$

- Since $\bar{X}\sim \mathrm{N}(\mu,\sigma^2/n)$, we have $n(\bar{X}-\mu)^2/\sigma^2\sim\chi^2(1)$
- It lacks showing that in the context of a normal population $\bar{X}$ e $S^{'2}$ are independent — and we shall prove this statement shortly — to conclude that

$$\frac{(n-1)S^{'2}}{\sigma^2}=\frac{nS^2}{\sigma^2}=\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2}\sim\chi^2(n-1)$$

Example 2.10 *Consider a normal population from which we have extracted a random sample of size 25. Suppose that we want to compute the probability that the ration between the bias-corrected sample variance and the population variance is between 0.79 and 1.18:*

$$
\begin{aligned}
P\left(0.79 < \frac{S'^2}{\sigma^2} < 1.18\right) &= P\left(18.96 < \frac{(n-1)S'^2}{\sigma^2} < 28.32\right) \\
&= \texttt{pchisq}(28.32, 24) - \texttt{pchisq}(18.96, 24) \\
&= 0.5073
\end{aligned}
$$

Student ratio

- When the population variance is unknown, the statement

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

is not very useful in practice

- In that situation, we have the "Student" ratio:

$$\frac{\bar{X}-\mu}{S'/\sqrt{n}} = \frac{\bar{X}-\mu}{S/\sqrt{n-1}} \sim t(n-1)$$

- We know that $\bar{X}$ e $S^{'2}$ are independent; note that

$$\frac{\bar{X}-\mu}{S'/\sqrt{n}} = \frac{\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{(n-1)S'^2}{\sigma^2}\dfrac{1}{n-1}}} = \frac{U}{\sqrt{V/(n-1)}}$$

where $U \sim \mathrm{N}(0,1)$ is independent of $V \sim \chi^2(n-1)$. The result is immediate.

- Recall that $t(n) \xrightarrow{d} \mathrm{N}(0,1)$: for large enough samples, $S^{'2}$ and σ^2 are very close...

Two normal populations

- $X_1 \sim \mathrm{N}(\mu_1, \sigma_1^2)$, $X_2 \sim \mathrm{N}(\mu_2, \sigma_2^2)$
- Two random samples, mutually independent, of size m and n respectively: $(X_{11}, \ldots, X_{1m})$ and $(X_{21}, \ldots, X_{2n})$

Difference of the sample means

- $\bar{X}_1 = \frac{1}{m} \sum_{i=1}^{m} X_{1i}$; $\bar{X}_2 = \frac{1}{n} \sum_{j=1}^{n} X_{2j}$
- It's easy to conclude that

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{m} + \dfrac{\sigma_2^2}{n}}} \sim \mathrm{N}(0, 1)$$

- The previous result is of limited use if the population variances are not known

- When the variance, although unknown, can be assumed as equal, we can resort to another result to make inference about $\mu_1 - \mu_2$: if $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then

$$T = \frac{\dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{1}{m} + \dfrac{1}{n}}}}{\sqrt{\dfrac{(m-1)S_1^{'2} + (n-1)S_2^{'2}}{m+n-2}}} \sim t(m+n-2)$$

since in this case

$$U = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\dfrac{1}{m} + \dfrac{1}{n}}} \sim \mathrm{N}(0,1)$$

and

$$V = \frac{(m-1)S_1^{'2} + (n-1)S_2^{'2}}{\sigma^2} \sim \chi^2(m+n-2)$$

are independent and $T = U/\sqrt{V/(m+n-2)}$.

- When the variances are unknown and different inferences about $\mu_1 - \mu_2$ become more complicated.
  - When the sample sizes are large, Slutsky's theorem allows us to replace the population variances by the sample variances and obtain the same distribution in the limit
  - for small sample sizes: Welch's approximation

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1'^2}{m} + \frac{S_2'^2}{n}}} \overset{a}{\sim} t(\nu)$$

where ν is the largest integer that does not exceed

$$\frac{\left(\frac{s_1'^2}{m} + \frac{s_2'^2}{n}\right)^2}{\frac{1}{m-1}\left(\frac{s_1'^2}{m}\right)^2 + \frac{1}{n-1}\left(\frac{s_2'^2}{n}\right)^2}$$

Two sample variances

- The two samples being independent, the random variables

$$U = \frac{(m-1)S_1'^2}{\sigma_1^2} \sim \chi^2(m-1)$$
$$V = \frac{(n-1)S_2'^2}{\sigma_2^2} \sim \chi^2(n-1)$$

are independent, hence

$$F = \frac{U/(m-1)}{V/(n-1)} = \frac{S_1'^2}{S_2'^2}\ \frac{\sigma_2^2}{\sigma_1^2} \sim F(m-1, n-1)$$

- In particular, when $\sigma_1^2 = \sigma_2^2$, we have

$$\frac{S_1'^2}{S_2'^2} \sim F(m-1, n-1)$$

Bernoulli population

- Two types of individuals in the population: the ones who possess a certain attribute and the ones who don't
- In what follows, let $(X_1, \dots, X_n)$ be an iid random sample of size n from a $\mathrm{B}(1, \theta)$ population.
- It is useful to establish the sampling distribution of two statistics: $T = \sum_{i=1}^{n} X_i$ and $\bar{X} = T/n$ — the number of individuals in the sample who possess the attribute and the proportion of individuals in the sample who possess the attribute.
- Clearly, $T \sim \mathrm{B}(n, \theta)$, hence

$$P(T = t) = \binom{n}{t} \theta^t (1 - \theta)^{n-t}, \; t = 0, \dots, n$$

$$P(\bar{X} = z) = \binom{n}{nz} \theta^{nz} (1 - \theta)^{n-nz}, \; z = 0/n, 1/n, \dots, n/n \; .$$

- Large sample approximations: De Moivre-Laplace theorem and law of rare events

- De Moivre-Laplace:

$$\frac{T - n\theta}{\sqrt{n\theta(1-\theta)}} \xrightarrow{d} \mathrm{N}(0,1) \qquad\qquad \frac{\bar{X} - \theta}{\sqrt{\theta(1-\theta)/n}} \xrightarrow{d} \mathrm{N}(0,1)$$

- Empirical rule: use when $n > 20$, $n\theta \geq 5$ and $n\theta(1-\theta) \geq 5$, $0.1 < \theta < 0.9$, together with the so-called continuity correction: with $a < b$, $a, b = 0, 1, \ldots, n$

$$P(a \leq T \leq b) \approx \Phi\left(\frac{b + 1/2 - n\theta}{\sqrt{n\theta(1-\theta)}}\right) - \Phi\left(\frac{a - 1/2 - n\theta}{\sqrt{n\theta(1-\theta)}}\right)$$

- The coefficient of symmetry of $\mathrm{B}(1,\theta)$ is $\gamma_1 = (1-2\theta)/\sqrt{\theta(1-\theta)}$ hence the farther from $1/2$ is θ the larger n needs to be

- Law of rare events

$$T \overset{a}{\sim} \mathrm{Po}(n\theta)$$

- Rule of thumb: use for $n > 20$ when $\theta \notin (0.1, 0.9)$ and $n\theta < 5$
- These approximations are nowadays useful for analytical purposes and for added insight. To compute the actual probabilities they are nowadays unnecessary.

Example 2.11 *Suppose a bank classifies its clients as "bad" if they have missed one or more credit card payments in the last two years. Suppose also that the proportion of "bad" clients ($X = 1$) is 0.05 for clients of the Lisbon area. What is the probability of obtaining more that 10% of "bad" clients in a random sample of: (a) 10 clients; (b) 50 clients; (c) 400 clients?* ■

Letting $\bar{X}$ denote the proportion of "bad" clients in the random sample, we need to compute $P(\bar{X} \geq 0.1)$

(a) Small sample

$$P(\bar{X} \geq 0.1) = P(T \geq 10 \times 0.1) = 1 - P(T = 0) = 1 - (1 - 0.05)^{10} = 0.4013$$

(b) $n = 50 > 20$, $\theta = 0.05 < 0.1$, $n\theta = 2.5 < 5$: use law of rare events

$P(\bar{X} \geq 0.1) = P(T \geq 5) = 1 - P(T \geq 4) \approx 1 - \texttt{ppois(4,50*0.05)} = 0.1088$

"exact" value: $1 - \texttt{pbinom(4,50,0.05)} = 0.1036$

- $n = 400 > 20$, $\theta = 0.05 < 0.01$, $n\theta = 20 \geq 5$: use normal approximation

Without continuity correction

$$P(\bar{X} \geq 0.1) \approx 1 - \Phi\left[(40 - 20)/\sqrt{400 \times 0.05 \times (1 - 0.05)}\right] = 2.23 \times 10^{-6}$$

With continuity correction

$$P(\bar{X} \geq 0.1) \approx 1 - \Phi\left[(40 - 1/2 - 20)/\sqrt{400 \times 0.05 \times (1 - 0.05)}\right] = 3.84 \times 10^{-6}$$

Law of rare events: $P(\bar{X} \geq 0.1) = 1 - \texttt{ppois(39,400*0.05)} = 5.32 \times 10^{-5}$
"Exact" value: $P(\bar{X} \geq 0.1) = 1 - \texttt{pbinom(39,400,0.05)} = 3.15 \times 10^{-5}$

Two Bernoulli populations

- Two Bernoulli populations with success probabilities θ_1 and θ_2.
- We want to compare θ_1 and θ_2 (for instance, the success rates for patients treated with drugs A and B)
- $\theta_1 - \theta_2$ will be unknown; we want to make inference about this quantity through the statistic $\bar{X}_1 - \bar{X}_2$, the difference between the sample proportions in two independent samples:
  - $(X_{11}, \ldots, X_{1m}) \Rightarrow \bar{X}_1 = \sum_{i=1}^{m} X_{1i}/m$
  - $(X_{21}, \ldots, X_{2n}) \Rightarrow \bar{X}_2 = \sum_{j=1}^{n} X_{2j}/n$
- Sampling distribution of $\bar{X}_1 - \bar{X}_2$?

- There are no exact results that are tractable
- Asymptotic distribution: by the De Moivre-Laplace, we have

$$\frac{\bar{X}_1 - \theta_1}{\sqrt{\theta_1(1-\theta_1)/m}} \xrightarrow{d} \mathrm{N}(0,1) \qquad \frac{\bar{X}_2 - \theta_2}{\sqrt{\theta_2(1-\theta_2)/n}} \xrightarrow{d} \mathrm{N}(0,1)$$

and using the independence we can show that (HW problem)

$$\frac{\bar{X}_1 - \bar{X}_2 - (\theta_1 - \theta_2)}{\sqrt{\dfrac{\theta_1(1-\theta_1)}{m} + \dfrac{\theta_2(1-\theta_2)}{n}}} \xrightarrow{d} \mathrm{N}(0,1)$$

Example 2.12 *(Cont'd) Suppose that the proportion of "bad" clients in the Oporto region is 0.06. If we obtain a sample of size 400 in the Lisbon region and of size 500 in the Oporto region, what is the chance of observing a larger proportion of "bad" clients in the Lisbon sample that in the Oporto sample?* ■

With $\theta_1 = 0.05$, $\theta_2 = 0.06$, $m = 400$, $n = 500$,

$$P(\bar{X}_1 - \bar{X}_2 > 0) = P\left(\frac{\bar{X}_1 - \bar{X}_2 - (\theta_1 - \theta_2)}{\sqrt{\dfrac{\theta_1(1-\theta_1)}{m} + \dfrac{\theta_2(1-\theta_2)}{n}}} > \frac{0 - (\theta_1 - \theta_2)}{\sqrt{\dfrac{\theta_1(1-\theta_1)}{m} + \dfrac{\theta_2(1-\theta_2)}{n}}}\right)$$
$$\approx 1 - \Phi(0.66) \approx 0.2546$$

This result shows that care must be taken when extrapolating conclusions from the samples to the whole universe.

Other populations: the gamma example

- If $X \sim \mathrm{G}(\alpha, \lambda)$, then (for $\alpha, \lambda > 0$)

$$f(x \mid \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\lambda x) \ , \quad x > 0$$

- If $\alpha \in \mathbb{N}$, it is also known as the Erlang distribution; $\alpha = 1$ results in the $\mathrm{Ex}(\lambda)$ distribution; $\mathrm{G}(n/2, 1/2) = \chi^2(n)$

- α is the shape parameter; λ is the rate. Sometimes this distribution is parametrize in terms of $\beta = 1/\lambda$ — the scale parameter:

```
dgamma(x, shape, rate = 1, scale = 1/rate, log = FALSE)
```

- If $X_1 \sim \mathrm{G}(\alpha_1, \lambda)$ is independent of $X_2 \sim \mathrm{G}(\alpha_2, \lambda)$ then

$$X_1 + X_2 \sim \mathrm{G}(\alpha_1 + \alpha_2, \lambda)$$

- If $c > 0$ and $X \sim \mathrm{G}(\alpha, \lambda)$, then

$$cX \sim \mathrm{G}(\alpha, \lambda/c)$$

- If $X \sim \mathrm{G}(\alpha, \lambda)$, then

$$2\lambda X \sim \mathrm{G}(\alpha, 1/2) = \chi^2(2\alpha)$$

- The following result is often used:

$$X \sim \chi^2(n) \Rightarrow \sqrt{2X} - \sqrt{2n-1} \xrightarrow{d} \mathrm{N}(0,1)$$

How do we show this?

- Let $X_1, \ldots, X_n$ be a random sample of size n from a $\mathrm{G}(\alpha, \lambda)$ population.
- Then

$$\sum_{i=1}^{n} X_i \sim \mathrm{G}(n\alpha, \lambda) \Leftrightarrow \bar{X} \sim \mathrm{G}(n\alpha, n\lambda) \Leftrightarrow 2n\lambda\bar{X} \sim \chi^2(2n\alpha)$$