STATISTICAL INFERENCE

- A Quick Guide -

Huan Q. Bui

Colby College

PHYSICS & MATHEMATICS
Statistics

Class of 2021

February 9, 2020

Preface

Greetings,

This guide is based on SC482: Statistical Inference, taught by Professor Liam O'Brien. The guide consists of lecture notes and material from *Introduction to Mathematical Statistics, 8th edition* by Hogg, McKean, and Craig. A majority of the text will be reading notes and solutions to selected problems.

As this is intended only to be a reference source, I might not be as meticulous with my explanations as I have been in some other guides.

Enjoy!

Contents

Part 1

Special Distributions

1.1 The Binomial and Related Distributions

If we let the random variable X equal the number of observed successes in n independent Bernoulli trials, each with success probability of p, then X follows the binomial distribution.s

A binomial pmf is given by

$$\boxed{p(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0, 1, 2, \dots \\ 0, \quad \text{else} \end{cases}} \tag{1.1}$$

Using the binomial expansion formula, we can easily check that

$$\boxed{\sum_x p(x) = 1} \tag{1.2}$$

The mgf of a binomial distribution is obtained by:

$$\boxed{M_{\text{bin}}(t) = E[e^{tx}] = \sum_x e^{tx} p(x) = \left[(1-p) + pe^t\right]^n \forall t \in \mathbb{R}} \tag{1.3}$$

With this, we can find the mean and variance for $p(x)$:

$$\boxed{\mu = M'(0) = n, \quad \sigma^2 = M''(0) = np(1-p)} \tag{1.4}$$

Theorem: Let $X_1, X_2, \dots, X_m$ be independent binomial random variables such that $X_i \sim \text{bin}(n_i, p), i = 1, 2, \dots, m$. Then

$$\boxed{Y = \sum_{i=1}^m X_i \sim \text{bin}\left(\sum_{i=1}^m n_i, p\right)} \tag{1.5}$$

Proof: We prove this via the mgf for Y. By independence, we have that

$$M_Y(t) = \prod_{i=1}^m (1 - p + pe^t)^{n_i} = (1 - p + pe^t)^{\sum_{i=1}^m n_i} \tag{1.6}$$

The mgf completely determines the distribution which Y follows, so we're done. □

1.1.1 Negative Binomial & Geometric Distribution

Consider a sequence of independent Bernoulli trials with constant probability p of success. The random variable Y which denotes the total number of failures in this sequence before the rth success follows the negative binomial distribution.

A negative binomial pmf is given by

$$\boxed{p_Y(t) = \begin{cases} \binom{(y+r-1)}{r-1} p^r (1-p)^y & y = 0, 1, 2, \dots \\ 0, \quad \text{else} \end{cases}} \tag{1.7}$$

The mgf of this distribution is

$$\boxed{M(t) = p^r [1 - (1-p)e^t]^{-r}} \tag{1.8}$$

When $r = 1$, Y follows the geometric distribution, whose pmf is given by

$$\boxed{p_Y(y) = p(1-p)^y, \quad y = 0, 1, 2, \dots} \tag{1.9}$$

The mgf of this distribution is

$$\boxed{M(t) = p[1 - (1-p)e^t]^{-1}} \tag{1.10}$$

1.2 Multinomial Distribution

We won't worry about this for now.

1.3 Hypergeometric Distribution

We won't worry about this for now.

1.4 The Poisson Distribution

The Poisson distribution gives the probability of observing x occurrences of some rare events characterized by rate $\lambda > 0$. The pmf is given by

$$\boxed{p(x) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!}, & x = 0, 1, 2, \dots \\ 0, \quad \text{else} \end{cases}} \tag{1.11}$$

We say a random parameter with the pmf of the form of $p(x)$ follows the Poisson distribution with parameter λ.

The mgf of a Poisson distribution is given by

$$\boxed{M(t) = e^{-\lambda(e^t - 1)}} \tag{1.12}$$

From here, we can find the mean and variance:

$$\boxed{\mu = M'(0) = \lambda, \quad \sigma^2 = M''(0) = \lambda} \tag{1.13}$$

Theorem: If $X_1, \ldots, X_n$ are independent random variables, each $X_i \sim \text{Poi}(\lambda_i)$, then

$$\boxed{Y = \sum_{i=1}^{n} X_i \sim \text{Poi}\left(\sum_{i=1}^{n} \lambda_i\right)} \tag{1.14}$$

Proof: We once again prove this via the mgf of Y:

$$M_Y(t) = \prod_{i=1}^{n} e^{\lambda_i(e^t-1)} = e^{\sum_{i=1}^{n} \lambda_i(e^t-1)} \tag{1.15}$$

□

1.5 The Γ, χ^2, β distributions

The gamma function of $\alpha > 0$ is given by

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y}\, dy, \tag{1.16}$$

which gives $\Gamma(1) = 1$ and $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$.

1.5.1 The Γ and exponential distribution

A continuous random variable $X \sim \Gamma(\alpha, \beta)$ where $\alpha > 0$ and $\beta > 0$ whenever its pdf is

$$\boxed{f(x) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, & 0 < x < \infty \\ 0, \quad \text{else} \end{cases}} \tag{1.17}$$

The mgf for X is obtained via the change of variable $y = x(1 - \beta t)/\beta$, where $t < 1/\beta$:

$$\boxed{M(t) = \int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x(1-\beta t)/\beta}\, dx = \frac{1}{(1-\beta t)^\alpha}} \tag{1.18}$$

From here, we can find the mean and variance:

$$\boxed{\mu = M'(0) = \alpha\beta, \quad \sigma^2 = \alpha\beta^2} \tag{1.19}$$

The $\Gamma(1, \beta)$ distribution is a special case, and it is called the **exponential distribution** with parameter $1/\beta$.

Theorem: Let $X_1, \ldots, X_n$ be independent random variables, with $X_i \sim \Gamma(\alpha_i, \beta)$. Then

$$\boxed{Y = \sum_{i=1}^{n} X_i \sim \Gamma\left(\sum_{i=1}^{n} \alpha_i, \beta\right)} \tag{1.20}$$

Proof: Can you guess via which device we prove the statement above? □

1.5.2 The χ^2 distribution

The χ^2 distribution is a special case of the gamma distribution where $\alpha = r/2, r \in \mathbb{N}^*$ and $\beta = 2$. If a continuous r.v. $X \sim \chi^2(r)$ then its pdf is

$$\boxed{f(x) = \begin{cases} \frac{1}{\Gamma(r/2)2^{r/2}} x^{r/2-1} e^{-x/2}, & 0 < x < \infty \\ 0, & \text{else} \end{cases}} \tag{1.21}$$

Its mgf is

$$\boxed{M(t) = (1-2t)^{-r/2}, \quad t < \frac{1}{2}} \tag{1.22}$$

Theorem: Let $X \sim \chi^2(r)$ and $k > -r/2$ be given. Then $E[X^k]$ exists and is given by

$$\boxed{E[X^k] = \frac{2^k \Gamma(r/2 + k)}{\Gamma(r/2)}} \tag{1.23}$$

Proof: is proof is purey computational and is left to the reader. □

From here, we note that all moments of the χ^2 distribution exist.

Theorem: Let $X_1, \ldots, X_n$ be r.v. with $X_i \sim \chi^2(r_i)$. Then

$$\boxed{Y = \sum_{i=1}^{n} X_i \sim \chi^2 \left(\sum_{i=1}^{n} r_i \right)} \tag{1.24}$$

Proof: we once again find the mgf for Y. □

1.5.3 The β distribution

The β distribution differs from the other continuous ones we've discussed so far because its support are bounded intervals.

I will skip most of the details here, except mentioning that we can derive the beta distribution from the a pair of independent Γ random variables. Suppose $Y = X_1/(X_1 + X_2)$ where $X_i \sim \Gamma(\alpha, \beta)$ then the pdf of Y is that of the beta distribution:

$$\boxed{g(y) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, & 0 < y < 1 \\ 0, & \text{else} \end{cases}} \tag{1.25}$$

The mean and variance of Y are

$$\boxed{\mu = \frac{\alpha}{\alpha+\beta}, \quad \sigma^2 = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}} \tag{1.26}$$

1.6 The Normal distribution

I have dedicated a large chunk in the QFT notes to evaluating Gaussian integrals, so I won't go into that here.

$X \sim \mathcal{N}(\mu, \sigma^2)$ whenever its pdf is

$$\boxed{f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right), \quad -\infty < x < \infty} \tag{1.27}$$

where μ and σ^2 are the mean and variance of X, respectively.

The mgf of X is can be obtained via the substitution $X = \sigma Z + \mu$:

$$\boxed{M(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)} \tag{1.28}$$

We note the following correspondence for $X = \sigma Z + \mu$:

$$X \sim \mathcal{N}(\mu, \sigma^2) \iff Z \sim \mathcal{N}(0, 1) \tag{1.29}$$

Theorem: $X \sim \mathcal{N}(\mu, \sigma^2) \implies V = (X-\mu)^2/\sigma^2 \sim \chi^2(1)$, i.e. a standardized, squared normal follows a chi-square distribution.

Proof: The proof isn't too hard. Let us write V as W^2 and so $W \sim \mathcal{N}(0, 1)$. We consider the cdf $G(v)$ for V, with $v \geq 0$:

$$G(v) = P(W^2 \leq v) = P(-\sqrt{v} \leq W \leq \sqrt{v}) = 2\int_0^{\sqrt{v}} \frac{1}{\sqrt{2\pi}} e^{-w^2/2}\, dw \tag{1.30}$$

with $G(v) = 0$ whenever $v < 0$. From here, we can see that the pdf for v, under the change of notation $w \to \sqrt{y}$, is

$$g(v) = G'(v) = \frac{d}{dv}\left\{\int_0^v \frac{1}{\sqrt{2\pi}\sqrt{y}} e^{-y/2}\, dy\right\}, \quad 0 \geq v \tag{1.31}$$

or 0 otherwise. This means

$$\boxed{g(v) = \begin{cases} \frac{1}{\sqrt{\pi}\sqrt{2}} v^{1/2-1} e^{-v/2}, & 0 < v < \infty \\ 0, \quad \text{else} \end{cases}} \tag{1.32}$$

Using the fact that $\Gamma(1/2) = \sqrt{\pi}$ and by verifying that $g(v)$ integrates to unity we show $V \sim \chi^2(1)$. □

Theorem: Let $X_1, \ldots, X_n$ be independent r.v. with $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. Then for constants $a_1, \ldots, a_n$

$$\boxed{Y = \sum_{i=1}^{n} a_i X_i \sim \mathcal{N}\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)} \tag{1.33}$$

Proof: We once again prove this kind of theorems via the mgf for Y:

$$\begin{aligned} M(t) &= \prod_{i=1}^{n} \exp\left(t a_i \mu_i + \frac{1}{2} a_i^2 \sigma_i^2\right) \\ &= \exp\left\{t \sum_{i=1}^{n} a_i \mu_i + \frac{1}{2} t^2 \sum_{i=1}^{n} a_i^2 \sigma_i^2\right\} \end{aligned} \tag{1.34}$$

which is the mgf for the normal with the corresponding mean and variance above. □

Corollary: Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. Then

$$\boxed{\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \sim \mathcal{N}\left(\mu, \sigma^2/n\right)} \tag{1.35}$$

Proof: the proof is left to the reader.

1.6.1 Contaminated Normal

We won't worry about this for now.

1.7 The Multivariate Normal

I'll just jump straight to the n-dimensional generalization. Evaluations of high-dimensional Gaussian integrals and moments can also be found in the QFT notes.

We say an n-dimensional random vector $\mathbf{X}$ has a multivariate normal distribution if its mgf is

$$\boxed{M_{\mathbf{X}}(t) = \exp\left(\mathbf{t}^\top \boldsymbol{\mu} + \frac{1}{2} \mathbf{t}^\top \boldsymbol{\Sigma} \mathbf{t}\right)} \tag{1.36}$$

for all $\mathbf{t} \in \mathbb{R}^n$, where $\boldsymbol{\Sigma}$ is a symmetric, positive semi-definite matrix and $\boldsymbol{\mu} \in \mathbb{R}^n$. For short, we say $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Theorem: Suppose $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is positive definite. Then

$$\boxed{\mathbf{Y} = (\mathbf{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi^2(1)} \tag{1.37}$$

Theorem: If $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then

$$\boxed{\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b} \sim \mathcal{N}_n(\mathbf{A}\boldsymbol{\mu} + \mathbf{b})} \tag{1.38}$$

Proof: The proof once again uses the mgf for $\mathbf{Y}$, but also some linear algebra manipulations. □

There are many other theorems and results related to this topic, but I won't go into them for now.

1.8 The t- and F-distributions

These two distributions are useful in certain problems in statistical inference.

1.8.1 The t-distribution

Suppose $W \sim \mathcal{N}(0,1)$ and $V \sim \chi^2(r)$ and that they are independent. Then the joint pdf of W and V, called $h(w,v)$, is the product of the pdf's of W and V:

$$h(w,v) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-w^2/2} \frac{1}{\Gamma(r/2)2^{r/2}} v^{r/2-1} e^{-v/2}, & w \in \mathbb{R}, v > 0 \\ 0, & \text{else} \end{cases} \tag{1.39}$$

Now we define a new variable $T = W/\sqrt{V/r}$ and consider the transformation:

$$t = \frac{w}{\sqrt{v/r}} \quad u = v \tag{1.40}$$

which bijectively maps the parameter space $(w,v) = \mathbb{R} \times \mathbb{R}^+$ to $(t,u) = \mathbb{R} \times \mathbb{R}^+$. The absolute value of the Jacobian of the transformation is given by

$$|J| = \left| \det \begin{pmatrix} \partial_t w & \partial_u w \\ \partial_t v & \partial_u v \end{pmatrix} \right| = \frac{\sqrt{u}}{\sqrt{r}}. \tag{1.41}$$

With this, the joint pdf of T and $U \equiv V$ is given by

$$g(t,u) = |J| h\left(\frac{t\sqrt{u}}{\sqrt{r}}, u\right) = \begin{cases} \frac{u^{r/2-1}}{\sqrt{2\pi}\Gamma(r/2)2^{r/2}} \exp\left[-\frac{u}{2}\left(1 + \frac{t^2}{r}\right)\right] \frac{\sqrt{u}}{\sqrt{r}}, & t \in \mathbb{R}, u \in \mathbb{R}^+ \\ 0, & \text{else} \end{cases} \tag{1.42}$$

By integrating out u we obtain the marginal pdf for T:

$$\begin{aligned} g_1(t) &= \int_{-\infty}^{\infty} g(t,u)\, du \\ &= \int_0^{\infty} \frac{u^{(r+1)/2-1}}{\sqrt{2\pi r}\Gamma(r/2)2^{r/2}} \exp\left[-\frac{u}{2}\left(1 + \frac{t^2}{r}\right)\right] du. \end{aligned} \tag{1.43}$$

Via the substitution $z = u[1 + (t^2/r)]/2$ we can evaluate the integral to find for $t \in \mathbb{R}$

$$\boxed{g_1(t) = \frac{\Gamma[(r+1)/2]}{\sqrt{\pi r}\Gamma(r/2)} \frac{1}{(1+t^2/r)^{(r+1/2)}}} \tag{1.44}$$

A r.v. T with this pdf is said to follow the t-distribution (or the Student's t-distribution) with r degrees of freedom. The t-distribution is symmetric about 0 and has a unique maximum at 0. As $r \in \infty$, the t-distribution converges to $\mathcal{N}(0,1)$.

The mean of $T \sim \text{Stu}(r)$ is zero. The variance can be found to be $\text{Var}(T) = E[T^2] = \frac{r}{r-2}$, so long as $r > 2$.

1.8.2 The F-distribution

Let $U \sim \chi^2(r-1), V \sim \chi^2(r_2)$ be given. Then the joint pdf of U and V is once again the product of their pdf's:

$$h(u,v) = \begin{cases} \frac{1}{\Gamma(r_1/2)\Gamma(r_2/2)2^{(r_1+r_2)/2}} u^{r_1/2-1} v^{r_2/2-1} e^{-(u+v)/2}, & u,v \in \mathbb{R}^+ \\ 0, \quad \text{else} \end{cases} \tag{1.45}$$

Define the new random variable

$$W = \frac{U/r_1}{V/r_2} \tag{1.46}$$

whose pdf $g_1(w)$ we are interested in finding. Consider the transformation

$$w = \frac{u/r_1}{v/r_2}, \quad z = v \tag{1.47}$$

which bijectively maps $(u,v) = \mathbb{R}^+ \times \mathbb{R}^+ \to (w,z) = [\mathbb{R}^+ \times \mathbb{R}^+]$. Like last time, the absolute value of the Jacobian can be found to be

$$|J| = \frac{r_1}{r_2} z. \tag{1.48}$$

The joint pdf $g(w,z)$ of the random variables W and $Z = V$ is obtained from by scaling $h(u,v)$ by $|J|$ and applying the variable transformation:

$$g(w,z) = \frac{1}{\Gamma(r_1/2)\Gamma(r_2/2)2^{\frac{r_1+r_2}{2}}} \left(\frac{r_1 z w}{r_2}\right)^{\frac{r_1-2}{2}} z^{\frac{r_2-2}{2}} \exp\left[-\frac{z}{2}\left(\frac{r_1 w}{r_2}+1\right)\right] \frac{r_1 z}{r_2} \tag{1.49}$$

so long as $(w, z) \in \mathbb{R}^+ \times \mathbb{R}^+$ and 0 otherwise. We then proceed to find the marginal pdf $g_1(w)$ of W by integrating out z. By considering the change of variables:

$$y = \frac{z}{2}\left(\frac{r_1 w}{r_2} + 1\right) \tag{1.50}$$

we can evaluate the integral and find the marginal pdf of W to be

$$g_1(w) = \begin{cases} \frac{\Gamma[(r_1+r_2)/2](r_1/r_2)^{r_1/2}}{\Gamma(r_1/2)\Gamma(r_2/2)} \frac{w^{r_1/2-1}}{(1+r_1 w/r_2)^{(r_1+r_2)/2}}, & w \in \mathbb{R}^+ \\ 0, \quad \text{else} \end{cases} \tag{1.51}$$

W, which is the ratio of two independent chi-quare variables U, V, is said to follow an F-distribution with degrees of freedom r_1 and r_2. We call the ratio $W = (U/r_1)/(V/r_2)$ the "F" ratio.

The mean of W is $E[F] = \frac{r_2}{r_2-2}$. When r_2 is large, $E[F] \to 1$.

1.8.3 The Student's Theorem

Here we will create the connection between the normal distribution and the t-distribution. This is an important result for the later topics on inference for normal random variables.

Theorem: Let $X_1, \ldots, X_n$ be iid r.v. with $X_i \sim \mathcal{N}(\mu, \sigma^2) \forall i$. Define the r.v.'s

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{1.52}$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2. \tag{1.53}$$

Then

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

(b) $\bar{X}$ and S^2 are independent.

(c) $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$.

(d) The variable $\bar{T} = (\bar{X} - \mu)/(S/\sqrt{n})$ follows the Student's t-distribution with $n-1$ degrees of freedom.

$$\bar{T} = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim \text{Stu}(n-1). \tag{1.54}$$

Proof: The proof is good, so I will reproduce it here. Because $X_i \sim \mathcal{N}(\mu, \sigma^2) \forall i$, $\mathbf{X} \sim \mathcal{N}_n\left(\mu\mathbf{1}, \sigma^2\mathbf{1}\right)$, where $\mathbf{1}$ denotes the n-vector whose components are all 1.

Now, consider $\mathbf{v}^\top = (1/n)\mathbf{1}^\top$. We see that $\bar{X} = \mathbf{v}^\top \mathbf{X}$. Define the random vector $\mathbf{Y} = (X_1 - \bar{X}, \ldots, X_2 - \bar{X})^\top$ and consider the (true) equality:

$$\mathbf{W} = \begin{pmatrix} \bar{X} \\ \mathbf{Y} \end{pmatrix} = \underbrace{\begin{pmatrix} \mathbf{v}^\top \\ \mathbb{I} - \mathbf{1}\mathbf{v}^\top \end{pmatrix}}_{\text{the transformation}} \mathbf{X} \tag{1.55}$$

which just restates our definitions nicely. We see that $\mathbf{W}$ is a result of a linear transformation of multivariate normal random vector, and so it follows that $\mathbf{W} \sim \mathcal{N}_{n+1}$ with mean

$$E[\mathbf{W}] = \begin{pmatrix} \mathbf{v}^\top \\ \mathbb{I} - \mathbf{1}\mathbf{v}^\top \end{pmatrix} \mu\mathbf{1} = \begin{pmatrix} \mu \\ \mathbf{0}_n \end{pmatrix} \tag{1.56}$$

and the covariance matrix

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{v}^\top \\ \mathbb{I} - \mathbf{1}\mathbf{v}^\top \end{pmatrix} \sigma^2 \mathbb{I} \begin{pmatrix} \mathbf{v}^\top \\ \mathbb{I} - \mathbf{1}\mathbf{v}^\top \end{pmatrix}^\top = \sigma^2 \begin{pmatrix} \frac{1}{n} & \mathbf{0}_n^\top \\ \mathbf{0}_n & \mathbb{I} - \mathbf{1}\mathbf{v}^\top \end{pmatrix} \tag{1.57}$$

From here, part (a) is proven. Next, observe that $\mathbf{\Sigma}$ is diagonal, and so all covariances are zero. This means $\bar{X}$ is independent of $\mathbf{Y}$. But because $S^2 = (n-1)^{-1}\mathbf{Y}^\top\mathbf{Y}$, $\bar{X}$ is independent of S^2 as well. So, (b) is proven.

Now, consider the r.v.

$$V = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \tag{1.58}$$

Each summand of V is a square of an $\mathcal{N}(0,1)$ r.v., and so each follows a $\chi^2(1)$. Because V is a sum of squares of n such $\chi^2(1)$'s, $V \sim \chi^2(n)$. Next, we can rewrite V as

$$V = \sum_{i=1}^{n} \left(\frac{(X_i - \bar{X}) + (\bar{X} - \mu)}{\sigma} \right)^2 = \frac{(n-1)S^2}{\sigma^2} + \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right)^2. \tag{1.59}$$

By (b), the summands in the last equation is are independent. The second term is a square of a $\mathcal{N}(0,1)$, so it follows a $\chi^2(1)$. Taking mgfs of both sides, we get

$$(1-2t)^{-n/2} = \underbrace{E[\exp\{t(n-1)S^2/\sigma^2\}]}_{M_{(c)}}(1-2t)^{-1/2}. \tag{1.60}$$

Solving for the mgf of $(n-1)S^2/\sigma^2$ we get part (c). Finally, writing T as

$$T = \frac{(\bar{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{(n-1)S^2/(\sigma^2(n-1))}} \tag{1.61}$$

and using (a)-(c) gives us (d). *Hint*: consider what distributions the numerator and denominator of T follow. □

1.9 Problems

3.6.4

(a) X has a standard normal distribution:

```
x=seq(-6,6,.01); plot(dnorm(x)~x)
```

(b) X has a t-distribution with 1 degree of freedom.

```
lines(dt(x,1)~x,lty=2)
```

(c) X has a t-distribution with 3 degrees of freedom.

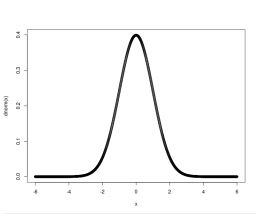
```
lines(dt(x,3)~x,lty=2)
```

(d) X has a t-distribution with 10 degrees of freedom.

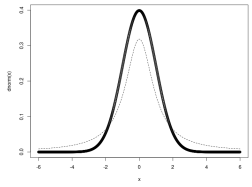
```
lines(dt(x,10)~x,lty=2)
```

(e) X has a t-distribution with 30 degrees of freedom.

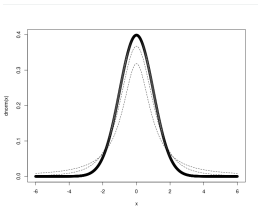
```
lines(dt(x,30)~x,lty=2)
```



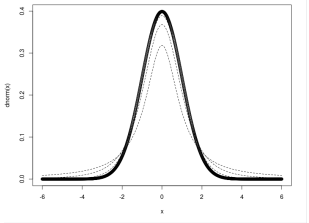
(a)



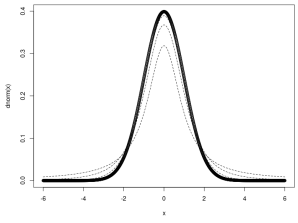
(b)



(c)



(d)



(e)

3.6.5

(a) $P(|X| \geq 2) = 2 \times [1 - P(X \leq 2)] = \mathbf{0.046}$.

```
> 2*(1 - pnorm(2))
[1] 0.04550026
```

(b) $P(|X| \geq 2) = 2 \times [1 - P(X \leq 2)] = \mathbf{0.295}$.

```
> 2*(1 - pt(2,1))
[1] 0.2951672
```

(c) $P(|X| \geq 2) = 2 \times [1 - P(X \leq 2)] = \mathbf{0.139}$.

```
> 2*(1 - pt(2,3))
[1] 0.139326
```

(d) $P(|X| \geq 2) = 2 \times [1 - P(X \leq 2)] = \mathbf{0.073}$.

```
> 2*(1 - pt(2,10))
[1] 0.07338803
```

(e) $P(|X| \geq 2) = 2 \times [1 - P(X \leq 2)] = \mathbf{0.055}$.

```
> 2*(1 - pt(2,30))
[1] 0.05462504
```

3.6.11: Let $T = W/\sqrt{V/r}$, where the independent variables $W \sim \mathcal{N}(0,1)$ and $V \sim \chi^2(r)$. Show that $T^2 \sim F(r_1 = 1, r_2 = r)$. *Hint*: What is the distribution of the numerator of T^2?

Solution: Let the independent random variables U, V be given, with $W \sim \mathcal{N}(0,1)$ and $U \sim \chi^2(r)$. The random variable T^2, where $T = W/\sqrt{V/r}$ is given by

$$T^2 = \left(\frac{W}{\sqrt{V/r}} \right)^2 = \frac{W^2}{V/r}. \tag{1.62}$$

Because $W \sim \mathcal{N}(0,1)$, we have that $W^2 \sim \chi^2(1)$ (by theorem). Now, T^2 has the form

$$T^2 = \frac{W^2}{V/r} = \frac{W^2/1}{V/r} \tag{1.63}$$

where 1 is the df of $\chi^2(1)$ which W follows, and r is the df of $\chi^2(r)$ which U follows. Thus, $T^2 \sim F(1, r)$, by the definition of the F-distribution. □

3.6.15: Let X_1, X_2 be iid with common distribution having the pdf

$$f(x) = \begin{cases} e^{-x}, & 0 < x < \infty \\ 0, & \text{else} \end{cases} \tag{1.64}$$

Show that $Z = X_1/X_2$ has an F-distribution.

Solution: It suffices to show that Z can be written as a ratio of two χ^2-distributed independent random variables. To this end, we can consider the mgf $M_X(t)$ of X_1, which is also identically that of X_2 since X_1, X_2 are iid:

$$M_X(t) = E[e^{tx}] = \int_0^\infty e^{tx} e^{-x}\, dx = (1-t)^{-1}. \tag{1.65}$$

However, this does not quite match the mgf for a $\chi^2(2)$. To circumvent this problem, we rewrite

$$Z = \frac{X_1}{X_2} = \frac{2X_1/2}{2X_2/2} = \frac{(X_1 + X_1)/2}{(X_2 + X_2)/2}, \tag{1.66}$$

as we expect $r = 2$. Let $Y_1 = X_1 + X_1$. Then we have trivially $Y_1 = 2X_1$, and so $|J| = 2$. With this, Y_1 has the pdf

$$\tilde{f}_Y(y) = |J| f(x) = 2f(x) = \begin{cases} 2e^{-y/2}, & 0 < y < \infty \\ 0, & \text{else} \end{cases} . \tag{1.67}$$

From here, we find the mgf of Y_1 to be

$$M_{Y_1}(t) = E[e^{ty}] = \frac{1}{2}\int_0^\infty e^{ty} e^{-y/2}\, dy = (1-2t)^{-1} = (1-2t)^{-2/2}, t < \frac{1}{2}. \tag{1.68}$$

By symmetry, $M_{Y_2}(t)$ is identically $M_{Y_1}(t)$, and both are the mgf for $\chi^2(r=2)$. Because each mgf uniquely determines a pdf, $Y_1, Y_2 \sim \chi^2(r=2)$ identically and independently (for each depends exclusively on X_1, X_2, respectively). Therefore,

$$Z = \frac{(X_1 + X_1)/2}{(X_2 + X_2)/2} = \frac{Y_1/2}{Y_2/2} \tag{1.69}$$

follows the F-distribution with degrees of freedom $r_1 = r_2 = 2$, by definition. □

3.6.16: Let X_1, X_2, X_3 be independent r.v. with $X_i \sim \chi^2(r_i)$.

(a) Show that $Y_1 = X_1/X_2$ and $Y_2 = X_1 + X_2$ are independent and that $Y_2 \sim \chi^2(r_1 + r_2)$.

(b) Deduce that

$$\frac{X_1/r_1}{X_2/r_2} \text{ and } \frac{X_3/r_3}{(X_1+X_2)/(r_1+r_2)} \tag{1.70}$$

are independent F-variables.

Solution:

(a) We consider the transformation

$$y_1 = u(x_1, x_2) = \frac{x_1}{x_2} \tag{1.71}$$

$$y_2 = v(x_1, x_2) = x_1 + x_2. \tag{1.72}$$

whose inverse is

$$\begin{aligned} x_1 &= \bar{u}(y_1, y_2) = \frac{y_1 y_2}{1+y_1} \\ x_2 &= \bar{v}(y_1, y_2) = \frac{y_2}{1+y_1}. \end{aligned} \tag{1.73}$$

The absolute value of the Jacobian is

$$|J| = \left| \det \begin{pmatrix} \partial_{y_1}\bar{u} & \partial_{y_2}\bar{u} \\ \partial_{y_1}\bar{v} & \partial_{y_2}\bar{v} \end{pmatrix} \right| = \frac{y_2}{(1+y_1)^2}, \tag{1.74}$$

which maps one-to-one from the space of X_1, X_2 $\mathbb{R}^+ \times \mathbb{R}^+$ onto the space of Y_1, Y_2 $\mathbb{R}^+ \times \mathbb{R}^+$. Since X_1, X_2 are independent, we consider the joint pdf of X_1, X_2:

$$h(x_1, x_2) = \begin{cases} \frac{x_1^{r_1/2-1} x_2^{r_2/2-1}}{\Gamma(r_1/2)\Gamma(r_2/2)2^{(r_1+r_2)/2}} e^{-(x_1+x_2)/2}, & 0 < x_1, x_2 < \infty \\ 0, \quad \text{else} \end{cases} \tag{1.75}$$

from which we can deduce the joint pdf for Y_1, Y_2:

$$\begin{aligned} \tilde{h}(y_1, y_2) &= |J| h\left(\frac{y_1 y_2}{1+y_1}, \frac{y_2}{1+y_1}\right) \\ &= \begin{cases} \frac{y_2 (y_1 y_2)^{r_1/2-1} y_2^{r_2/2-1} (1+y_1)^{-r_1/2-r_2/2+2}}{\cancel{(1+y_1)^2}\Gamma(r_1/2)\Gamma(r_2/2)2^{(r_1+r_2)/2}} e^{-y_2/2}, & 0 < y_1, y_2 < \infty \\ 0, \quad \text{else} \end{cases} \\ &= \begin{cases} \frac{y_2^{r_1/2+r_2/2-1} y_1^{r_1/2-1} (1+y_1)^{-r_1/2-r_2/2}}{\Gamma(r_1/2)\Gamma(r_2/2)2^{(r_1+r_2)/2}} e^{-y_2/2}, & 0 < y_1, y_2 < \infty \\ 0, \quad \text{else} \end{cases} \end{aligned} \tag{1.76}$$

Without further computation we see that $\tilde{h}(y_1, y_2)$ can be written as a product of two nonnegative functions of y_1 and y_2. In view of Theorem 2.4.1, Y_1 and Y_2 are independent. □

Next, we wish to show $Y_2 \sim \chi^2(X_1, X_2)$, to which end we find the marginal pdf $g_2(y_2)$ of Y_2:

$$\begin{aligned} g_2(y_2) &= \int_0^\infty \tilde{h}(y_1, y_2)\, dy_1 \\ &= \mathfrak{C} \int_0^\infty y_1^{r_1/2-1}(1+y_1)^{-r_1/2-r_2/2}\, dy_1 \\ &= \mathfrak{C} \frac{\Gamma(r_1/2)\Gamma(r_2/2)}{\Gamma[(r_1+r_2)/2]} \end{aligned} \tag{1.77}$$

where $\mathfrak{C}$ contains all the y_1-independent elements. From here, via simple back-substitution we obtain the marginal pdf for Y_2:

$$g_2(y_2) = \begin{cases} \frac{y_2^{(r_1+r_2)/2-1}}{\Gamma[(r_1+r_2)/2]2^{(r_1+r_2)/2}} e^{-y_2/2}, & 0 < y_2 < \infty \\ 0, \quad \text{else} \end{cases}, \tag{1.78}$$

i.e., $Y_2 \sim \chi^2(r_1 + r_2)$. □

Mathematica code:

```
In[20]:= Integrate[
x^(r1/2 - 1) (1 + x)^(-r1/2 - r2/2), {x, 0, Infinity}]

Out[20]= ConditionalExpression[(Gamma[r1/2] Gamma[r2/2])/
Gamma[(r1 + r2)/2], Re[r2] > 0 && Re[r1] > 0]
```

(b) By definition, because X_1, X_2 are independent random variables with $X_i \sim \chi^2(r_i)$,

$$\Omega = \frac{X_1/r_1}{X_2/r_2} \sim F(r_1, r_2). \tag{1.79}$$

Also, because $X_3 \sim \chi^2(r_3)$ and $(X_1 + X_2) \sim \chi^2(r_1 + r_2)$ (from (a)), we have

$$\Lambda = \frac{X_3/r_3}{(X_1+X_2)/(r_1+r_2)} \sim F(r_3, r_1+r_2) \tag{1.80}$$

as well. Furthermore, because

$$\Omega = \frac{X_1/r_1}{X_2/r_2} = \frac{r_2}{r_1} Y_1 \tag{1.81}$$

$$\Lambda = \frac{r_1+r_2}{r_3} \frac{X_3}{Y_2} \tag{1.82}$$

and because X_1, X_2, X_3 are independent, we have that Y_1, Y_2, X_3 are independent. Therefore, it is necessary that $\Omega \sim F(r_1, r_2)$ and $\Lambda \sim F(r_3, r_1 + r_2)$ are independent as well. □

Part 2

Elementary Statistical Inferences

2.1 Sampling & Statistics

In statistical inferences, our ignorance about the pdf/pmf of a random variable X can be classified in two ways:

- The pdf/pmf is unknown.
- The pdf/pmf is assumed/known but its parameter vector θ is not.

We consider the second class of classification for now.

Definition: If the random variables $X_1, X_2, \ldots, X_n$ are iid, then these random variables constitute a **random sample** of size n from the common distribution.

Definition: Let $X_1, \ldots, X_n$ denote a sample on a random variable X. Let $T = T(X_1, \ldots, X_n)$ be a function of the sample. Then T is called a **statistic**.

2.1.1 Point estimators

Definition: (*Unbiasedness*) Let $X_1, \ldots, X_n$ denote a sample on a random variable X with pdf $f(x;\theta)$, $\theta \in \Omega$. Let $T = T(X_1, \ldots, X_n)$ be a statistic. We say that T is an *unbiased* estimator of θ if $E[T] = \theta$.

We now introduce the concept of the **maximum likelihood estimator (mle)**. The information in the sample and the parameter θ are involved in the joint distribution of the random sample. We write this as

$$\boxed{L(\theta) = L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i;\theta).} \tag{2.1}$$

This is called the **likelihood function** of the random sample. A measure of the center of $L(\theta)$ seems to be an appropriate estimate of θ. We often use the value of θ at which $L(\theta)$ is maximized. If this value is unique, then it is called the **maximum likelihood estimator** (mle), denoted as $\hat{\theta}$:

$$\hat{\theta} = \text{Argmax} L(\theta). \tag{2.2}$$

We often work with the log of the likelihood in practice, which is the function $l(\theta) = \log(L(\theta))$. The logarithm is a strictly increasing function, so its maximum is obtained exactly when the maximum of $L(\theta)$ is obtained. In most models, the pdf and pmf are differentiable functions of θ, in which cases $\hat{\theta}$ solves the equation:

$$\boxed{\partial_\theta l(\theta) = 0} \tag{2.3}$$

This is equivalent to saying $\hat{\theta}$ maximizes $l(\theta)$. If θ is a vector of parameters, this results in a system of equations to be solved simultaneously. These equations are called the **estimating equations**, (EE).

2.1.2 Histogram estimates of pmfs and pdfs

Let $X_1, \ldots, X_n$ be a random sample on a random variable X with cdf $F(x)$. A histogram of the sample is an estimate of the pmf or pdf depending on whether X is discrete of continuous. We make no assumptions on the form of the distribution of X. In particular, we don't assume the parametric form of the distribution, hence the histogram is often called the **nonparametric** estimator.

The distribution of X is discrete

Assume X is a discrete r.v. with pmf $p(x)$. Consider a sample $X_1, \ldots, X_n$. Suppose $X \in \mathcal{D} = \{a_1, \ldots, a_n\}$, then intuitively the estimate of $p(a_j)$ is the relative frequency of a_j. More formally, for $j = 1, \ldots, m$ we define the statistic

$$I_j(X_i) = \begin{cases} 1 & X_i = a_j \\ 0 & X_i \neq a_j \end{cases} \tag{2.4}$$

Then the estimate of $p(a_j)$ is the average

$$\hat{p}(a_j) = \frac{1}{n} \sum_{i=1}^{n} I_j(X_i) \tag{2.5}$$

The estimators $\{\hat{p}(a_1), \ldots, \hat{p}(a_m)\}$ constitute the nonparametric estimate of the pmf $p(x)$. We note that $I_j(X_i)$ has a Bernoulli distribution with probability $p(a_j)$, and so

$$E[\hat{p}(a_j)] = \frac{1}{n} \sum_{i=1}^{n} E[I_j(X_i)] = \frac{1}{n} \sum_{i=1}^{n} p(a_j) = p(a_j), \tag{2.6}$$

which means $\hat{p}(a_j)$ is an *unbiased estimator* of $p(a_j)$.

Now, suppose that the space of X is infinite, i.e, $\mathcal{D} = \{a_1, \ldots, \}$ then in practice we select a value, say a_m, and make the groupings

$$\{a_1\}, \{a_2\}, \ldots, \{a_m\}, \tilde{a}_{m+1} = \{a_{m+1}, \ldots\} \tag{2.7}$$

Let $\hat{p}(\tilde{a}_{m+1})$ be the proportion of the sample items that are greater than or equal to a_{m+1}. Then the estimates $\{\hat{p}(a_1), \ldots \hat{p}(a_{m+1})\}$ form our estimate of $p(x)$. To merge groups, the rule of thumb is to select m so that the frequency of the category a_m exceeds twice the combined frequencies of the categories $a_{m+1}, a_{m+2}, \ldots$

A histogram is a *barplot* pf $\hat{p}(a_j)$ versus a_j. When a_j contains no ordinal information (e.g. hair colors, etc) then such histograms consist of nonabutting bars and are called *bar charts*. When the space $\mathcal{D}$ is ordinal, then the histograms is an abutting bar chart plotted in the natural order of the a_j's.

The distribution of X is continuous

Assume X is a continuous r.v. with pdf $f(x)$. Consider a sample $X_1, \ldots, X_n$. We first sketch an estimate for this pdf at a specified value of x. For a given $h > 0$, we consider the interval $(x-h, x+h)$. By MVT, we have for $\xi, |x - \xi| < h$:

$$P(|X - x| < h) = \int_{x-h}^{x+h} f(t)\, dt \approx 2hf(x). \tag{2.8}$$

The LHS is the porprotion of the sample items that fall in the interval $(x - h, x + h)$. This suggests the use of the estimate of $f(x)$ at a given x:

$$\hat{f}(x) = \frac{\#\{|X_i - x| < h\}}{2hn}. \tag{2.9}$$

More formally, the indicator statistic is, for $i = 1, \ldots, n$

$$I_i(x) = \begin{cases} 1 & x - h < X_i < x + h \\ - & \text{else} \end{cases}, \tag{2.10}$$

from which we obtain the nonparametric estimator of $f(x)$:

$$\hat{f}(x) = \frac{1}{2hn} \sum_{i=1}^{n} I_i(x). \tag{2.11}$$

Since the sample items are iid:

$$E[\hat{f}(x)] = \frac{1}{2hn} nf(\xi)2h = f(\xi) \to f(x) \quad \text{as } h \to 0. \tag{2.12}$$

Therefore $\hat{f}(x)$ is *approximately* (as opposed to *exact* in the discrete case) an unbiased estimator of $f(x)$. I_i is called the **rectangular kernel** with **bandwidth** $2h$.

Provided realized values $x_1, \ldots, x_n$ of the random sample of X with pdf $f(x)$, there are many ways to obtained a histogram estimate of $f(x)$. First, select an integer m, an $h > 0$, and a value $a < \min(x_i)$, so that the m intervals cover the range of the sample. These intervals form our classes. Let $A_j = (a + (2j - 3)h, a + (2j - 1)h]$ for $j = 1, \ldots, m$. Let $\hat{f}_h(x)$ denote our histogram estimate. For $a - h < x \leq (2m - 1)h$, x is in one and only one A_j. Then for $x \in A_j$, we define

$$\hat{f}_h(x) = \frac{\#\{x_i \in A_j\}}{2hn} \geq 0. \tag{2.13}$$

We see that

$$\begin{aligned}\int_{\infty}^{\infty} \hat{f}_h(x)\,dx &= \int_{a-h}^{a+(2m-1)h} \hat{f}_h(x)\,dx \\ &= \sum_{j=1}^{m} \int_{A_j} \frac{\#\{x_i \in A_j\}}{2hn}\,dx \\ &= \frac{1}{2hn} \sum_{j=1}^{m} \#\{x_i \in A_j\}[h(2j-1-2j+3)] \\ &= \frac{2h}{2hn} n \\ &= 1. \end{aligned} \tag{2.14}$$

So $\hat{f}_h(x)$ satisfies the properties of a pdf.

2.2 Confidence Intervals

Definition: Let $X_1, \ldots, X_n$ be a sample on a r.v. X which has the pdf $f(x;\theta), \theta \in \Omega$. Let $0 < \alpha < 1$ be specified. Let $L = L(X_1, \ldots, X_n)$ and $U = U(X_1, \ldots, X_n)$ be two statistics. We say that the interval (L, U) is a $(1-\alpha)100\%$ **confidence interval** for θ if

$$1 - \alpha \equiv P_\theta[\theta \in (L, U)]. \tag{2.15}$$

That is, the probability that the interval includes θ is $1-\alpha$, which is called the **confidence coefficient** or the **confidence level** of the interval.

Under normality, the confidence interval for μ is given by

$$\boxed{\left(\bar{x} - t_{\alpha/2,n-1} s/\sqrt{n}, \bar{x} + t_{\alpha/2,n-1} s/\sqrt{n}\right)} \tag{2.16}$$

where $t_{\alpha/2,n-1}$ is the upper $\alpha/2$ critical points of a t-distribution with $n-1$ df. This CI is referred to as the $(1-\alpha)100\%$ ***t*-interval** for μ. s is referred to as the **standard error** of $\bar{X}$.

The Central Limit Theorem: Let $X_1, \ldots, X_n$ denote the observations of a random sample from a distribution that has mean μ and finite variance σ^2. Then the distribution function of the r.v. $W_n = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ converges to Φ, the distribution function of the $\mathcal{N}(0,1)$ distribution, as n $to\infty$.

When the sample is large, the CI for μ can be given by

$$\boxed{\left(\bar{x} - z_{\alpha/2} s/\sqrt{n}, \bar{z} + z_{\alpha/a} s/\sqrt{n}\right)} \tag{2.17}$$

In general, for the same α, the t-CI is larger (and hence more conservative) than the z-CI. When σ is known, we replace s by σ.

The larger sample CI for p is given by

$$\boxed{\left(\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\right)} \tag{2.18}$$

where $\sqrt{\hat{p}(1-\hat{p})/n}$ is called the standard error of $\hat{p}$.

2.2.1 CI for difference in means

By independence of samples,

$$\text{Var}(\hat{\Delta}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \tag{2.19}$$

where $\hat{\Delta} = \bar{X} - \bar{Y}$. We can readily show that $\hat{\Delta}$ is an unbiased estimator of $\Delta = \mu_1 - \mu_2$. Let the sample variances

$$S_j^2 = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(X_i - \bar{X})^2 \tag{2.20}$$

be given. Then th random variable follows the $\mathcal{N}(0,1)$:

$$\boxed{Z = \frac{\hat{\Delta} - \Delta}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim \mathcal{N}(0,1)} \tag{2.21}$$

The approximate $(1-\alpha)100\%$ CI for $\Delta = \mu_1 - \mu_2$ is then given by

$$\left((\bar{x} - \bar{y}) - z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, (\bar{x} - \bar{y}) + z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right) \tag{2.22}$$

This is a large sample $(1-\alpha)100\%$ CI for $\mu_1 - \mu_2$.

Now, suppose $X \sim \mathcal{N}(\mu_1, \sigma^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma^2)$ (i.e. X and Y are normally distributed with the same variance) are independent. We want to show $\Delta \sim t$-distribution. We know that $\bar{X} \sim \mathcal{N}(\mu_1, \sigma^2/n_1)$ and $\bar{Y} \sim \mathcal{N}(\mu_2, \sigma^2/n_2)$, so it is true that

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}} \sim \mathcal{N}(0,1). \tag{2.23}$$

This quantity will later be the numerator of our T-statistic. Now, let

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \tag{2.24}$$

then S_p^2, the **pooled estimator** of σ^2, is also an unbiased estimator of σ^2. Because $(n_i-1)S_i^2/\sigma^2 \sim \chi^2(n-1)$, we have that $(n-2)S_p^2/\sigma^2 \sim \chi^2(n-2)$. And so

$$\boxed{T = \frac{\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sigma\sqrt{1/n_1+1/n_2}}}{\sqrt{(n-2)S_p^2/(n-2)\sigma^2}} = \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_p\sqrt{1/n_1+1/n-2}} \sim t_{n-2}} \tag{2.25}$$

From here, it is easy to work out the $(1-\alpha)100\%$ CI for $\mu_1-\mu_2$:

$$\boxed{\left((\bar{x}-\bar{y}) - t_{\alpha/2,n-2}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}, (\bar{x}-\bar{y}) + t_{\alpha/2,n-2}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)} \tag{2.26}$$

There is some difficulty when the unknown variances σ in the distributions of X and Y are not equal.

2.2.2 CI for difference in proportions

Our estimator of the difference in proportions p_1-p_2 is $\bar{X}-\bar{Y} \equiv \hat{p}_1-\hat{p}_2$ where $X \sim b(1,p_1)$ and $Y \sim b(1,p_2)$. Of course, we know that $\sigma_1^2 = p_1(1-p_1)$ and $\sigma_2^2 = p_2(1-p_2)$. From here, the approximate $(1-\alpha)100\%$ confidence interval for p_1-p_2 is

$$\boxed{(\hat{p}_1-\hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \tag{2.27}$$

2.3 Order Statistics

Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution of the continuous type having a pdf $f(x)$ that has support $S=(a,b)$, where $\infty \le a < b \le \infty$. Let $Y_1 < Y_2 < \cdots < Y_n$ represent $X_1, X_2, \ldots, X_n$ when the latter are arranged in ascending order of magnitude. We call $Y_i, i=1,2,...,n$, the ith order statistic of the random sample $X_1, X_2, \ldots, X_n$. We have a theorem which gives the joint pdf of $Y_1, Y_2, \ldots, Y_n$.

Theorem: The joint pdf of $Y_1, Y_2, \ldots, Y_n$ is given by

$$g(y_1, y_2, \ldots, y_n) = \begin{cases} n!f(y_1)\ldots f(y_n) & a < y_1 < \cdots < y_n < b \\ 0 \quad \text{else} \end{cases} . \tag{2.28}$$

Proof: The support of $X_1, \ldots, X_n$ can be partitioned into $n!$ mutually disjoint sets that map onto the support of the Y_i's. Obviously the Jacobian for each

transformation is either 1 or -1.

$$g(y_1, y_2, \ldots, y_n) = \sum_{i=1}^{n!} |J_i| f(y_1) \ldots f(y_n)$$
$$= \begin{cases} n! f(y_1) \ldots f(y_n) & a < y_1 \cdots < y_n < b \\ 0 \quad \text{else} \end{cases} \tag{2.29}$$

2.3.1 Quantiles

Let X be a random variable with a continuous cdf $F(x)$. For $0 < p < 1$, define the pth quantile of X to be $\xi_p = F^{-1}(p)$. Let $X_1, X_2, \ldots, X_n$ be a random sample from the distribution of X and let $Y_1 < Y_2 < \cdots < Y_n$ be the corresponding order statistics. Let k be the greatest integer less than or equal to $p(n+1)$. We next define an estimator ofξ_p after making the following observation. The area under the pdf $f(x)$ to the left of Y_k is $F(Y_k)$. The expected value of this area is

$$E[F(Y_k)] = \int_a^b F(y_k) g_k(y_k)\, dy_k \tag{2.30}$$

where $g_k(y_k)$ is the pdf of Y_k. Consider the transformation $z = F(y_k)$, then the integral becomes

$$E[F(Y_k)] = \int_0^1 \frac{n!}{(k-1)!(n-k)!} z^k (1-z)^{n-k}\, dz = \cdots = \frac{k}{n+1} \tag{2.31}$$

where we recognize the similarity between the integral and the integral of a beta pdf. So, on the average, there is $k/(n+1)$ of the total area to the left of Y_k. Because $p = k/(n+1)$, it seems reasonable to take Y_k as an estimator of the quantile ξ_p. Hence, we call Y_k the **pth sample quantile**. It is also called the **$100p$th percentile** of the sample.

A **five-number** summary of the data consists of the following five sample quantiles: the minimum (Y_1), the first quartile ($Y_{.25(n+1)}$), the median, the third quartile ($Y_{.75(n+1)}$), and the maximum (Y_n). For this section, we use the notation Q_1, Q_2, and Q_3 to denote, respectively, the first quartile, median, and third quartile of the sample.

The five-number summary is the basis for a useful and quick plot of the data. This is called a **boxplot** of the data. In the **box and whisker** plots, we also define a potential outlier. Let $h = 1.5(Q_3 - Q_1)$. The **lower/upper fence** is defined by $L/UF = Q_{1/3} \mp h$. Points lying outside the (LF, UF) interval are called **potential outliers**.

2.3.2 CI for quantiles

Let X be a continuous random variable with cdf $F(x)$. For $0 < p < 1$, define the $100p$th distribution percentile to be ξ_p, where $F(\xi_p) = p$. For a sample of

size n on X, let $Y_1 < Y_2 < \cdots < Y_n$ be the order statistics. Let $k = [(n+1)p]$. Then the $100p$th sample percentile Y_k is a point estimate of ξ_p.

Let $i < [(n+1)p] < j$, and consider the order statistics $Y_i < Y_j$ and the event $Y_i < \xi_p < Y_j$. The event $Y_i < \xi_p < Y_j$ is equivalent to obtaining between i (inclusive) and j (exclusive) successes in n independent trials. So,

$$P(Y_i < \xi_p < Y_j) = \sum_{w=i}^{j-1} \binom{n}{w} p^w (1-p)^{n-w}. \tag{2.32}$$

For the median, we denote $\xi_{1/2}$ the median of $F(x)$, i.e. $\xi_{1/2}$ solves $F(x) = 1/2$. Let Q_2 denote the sample median, which is a point estimator of $\xi_{1/2}$. Take $c_{\alpha/2}$ such that $P[S \leq c_{\alpha/2}] = \alpha/2$ where $S \sim b(n, 1/2)$. Then note also that $P[S \leq c_{\alpha/2}] = \alpha/2$. From here we have

$$P[Y_{c_{\alpha/2}} < \xi_{1/2} < Y_{n-c_{\alpha/2}}] = 1 - \alpha. \tag{2.33}$$

So, if $y_{\alpha/2+1}$ and $y_{n-\alpha/2}$ are the realized values of the order statistics $Y_{c_{\alpha/2}+1}$ and $Y_{n-c_{\alpha/2}}$ then the interval

$$\boxed{\left(y_{c_{\alpha/2}}, y_{n-c_{\alpha/2}}\right)} \tag{2.34}$$

is a $(1-\alpha)100\%$ confidence interval for $\xi_{1/2}$.

2.4 Introduction to Hypothesis Testing

Suppose a r.v. $X \sim f(x;\theta)$, where $\theta \in \Omega$. Suppose that $\theta \in \omega_0$ or $\theta \in \omega_1$ where ω_0 and ω_1 are disjoint subsets of Ω and $\omega_0 \cup \omega_1 = \Omega$. We label these hypotheses as

$$\begin{aligned} H_0 &: \theta \in \omega_0 \\ H_1 &: \theta \in \omega_1. \end{aligned} \tag{2.35}$$

H_0 is called the **null hypothesis**. H_1 is called the **alternative hypothesis**. Type I error occurs when we decide that $\theta \in \omega_1$ when in fact $\theta \in \omega_0$. Type II error occurs when we decide the opposite.

We require the **critical region**, C, to complete the testing structure for the general problem. Consider the r.v. X and the hypotheses given above. C is such that

$$\begin{aligned} &\text{Reject } H_0 \text{ if } (X_1, \ldots, X_n) \in C \\ &\text{Reject } H_1 \text{ if } (X_1, \ldots, X_n) \in C^c. \end{aligned} \tag{2.36}$$

Type I error occurs if H_0 is rejected when it is true. **Type II** error occurs if H_0 is retained when H_1 is true.

Definition: We say a critical region C is of **size** α if

$$\boxed{\alpha = \max_{\theta \in \omega_0} P_\theta[(X_1, \ldots, X_n)] \in C} \tag{2.37}$$

Over all critical regions of size α, we want to consider critical regions that have lower probabilities of Type II error, i.e., for $\theta \in \omega_1$, we want to maximize

$$1 - P_\theta[\text{Type II Error}] = P_\theta[(X_1, \ldots, X_n) \in C] \tag{2.38}$$

The probability on the right side of the equation above is called the **power** of the test at θ. It is the probability that the test detects the alternative when $\theta \in \omega_1$ is the true parameter. Minimizing Type II error requires maximizing the test power. The **power function** of C is

$$\gamma_C(\theta) = P_\theta[(X_1, \ldots, X_n) \in C]; \quad \theta \in \omega_1. \tag{2.39}$$

Given two critical regions C_1, C_2 both of size α. C_1 is better than C_2 if $\gamma_{C_1}(\theta) \geq \gamma_{C_2}(\theta)$ for all $\theta \in \omega_1$.

A **simple** hypothesis completely specifies the underlying distribution. A **composite** hypothesis can be composed of many simple hypotheses and hence do not completely specify the distribution. α is often referred to as the **significance level** of the test associated with that critical region.

2.5 Additional comments about statistical test

2.5.1 Observed Significance Level, p-value

Suppose $H_0 : \mu = \mu_0$ and $H_1 : \mu > \mu_0$, where μ_0 is maximized. Then we reject H_0 in favor H_1 if $\bar{X} \geq k$ where $\bar{X}$ is the sample mean. The p-value is the probability that under H_0, $\bar{X} \geq \bar{x}$:

$$\boxed{p - \text{value} = P_{H_0}(\bar{X} \geq \bar{x})} \tag{2.40}$$

If $\alpha > p$ then we reject H_0 in favor of H_1. Else, we fail to reject H_1.

2.6 Chi-Square Tests

2.7 The Method of Monte Carlo

2.7.1 Accept-Reject Generation Algorithm

2.8 Bootstrap Procedures

2.8.1 Percentile Bootstrap CI

2.8.2 Bootstrap Testing Procedures

2.9 Problems

4.1.1 Twenty motors were put on test under a high-temperature setting. The lifetimes in hours of the motors under these conditions are given below. Also, the data are in the file **lifetimemotor.rda** at the site listed in the Preface. Suppose we assume that the lifetime of a motor under these conditions, X, has a $\Gamma(1, \theta)$ distribution.

1	4	5	21	22	28	40	42	51	53
58	67	95	124	124	160	202	260	303	363

(a) Obtain a histogram of the data and overlay it with a density estimate, using the code **hist(x,pr=T); lines(density(x))** where the R vector**x** contains the data. Based on this plot, do you think that the $\Gamma(1, \theta)$ model is credible?

(b) Assuming a $\Gamma(1, \theta)$ model, obtain the maximum likelihood estimate $\hat{\theta}$ of θ and locate it on your histogram. Next overlay the pdf of a $\Gamma(1, \hat{\theta})$ distribution on the histogram. Use the R function **dgamma(x,shape=1,scale=$\hat{\theta}$)** to evaluate the pdf.

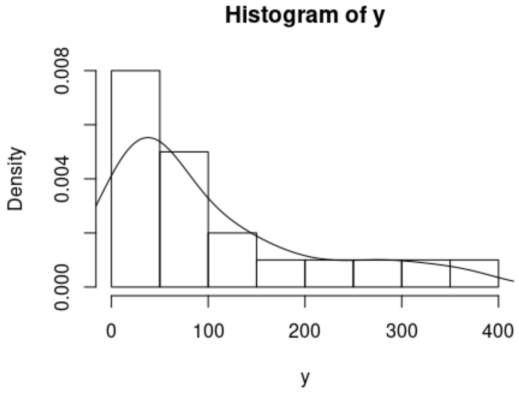
(c) Obtain the sample median of the data, which is an estimate of the median lifetime of a motor. What parameter is it estimating (i.e., determine the median of X)?

(d) Based on the mle, what is another estimate of the median of X?

Solution:

(a) For some reason R does not recognize the dataset as of numeric type. Because the dataset is small enough, I recoded and fed it by hand to the data vector y:

```
> lines(density(y))
> y <- c(1,4,5,21,22,28,40,42,51,53,58,67,
        95,124,124,160,202,260,303,363)
> hist(y,pr=T)
> lines(density(y))
```



The $\Gamma(1,\theta)$, or $\text{Exp}(\theta)$, model seems to be **credible** as far as the histogram is concerned. However, the overlaying density does not look like a $\Gamma(1,\theta)$. □

(b) Assuming the $\Gamma(1,\theta)$ model, then the pdf on the support $\mathbb{R}^+$ is given by

$$f(y) = \frac{1}{\theta} e^{-y/\theta}, \tag{2.41}$$

from which we obtain the logarithm of the likelihood function:

$$l(\theta) = \log\left(\prod_{i=1}^{n} \frac{1}{\theta} e^{-y_i/\theta}\right) = -n\log\theta - \frac{1}{\theta}\sum_{i=1}^{n} y_i. \tag{2.42}$$

The first partial derivative wrt θ is then

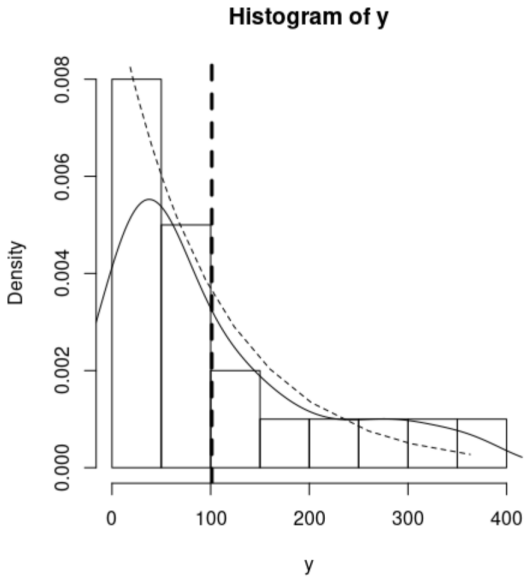
$$\partial_\theta l(\theta) = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} y_i. \tag{2.43}$$

Setting $\partial_\theta l(\theta) = 0$, we get (by inspection) that $l(\theta)$ is extremized iff $\theta = (1/n)\sum_{i=1}^{n} y_i = \bar{y}$. We also have that $\partial_{\theta\theta} < 0\,\forall\theta \in \mathbb{R}^+$, which means $l(\theta)$ is maximized globally at $\bar{y}$. From here, the statistic

$$\hat{\theta} = \bar{Y} = \mathbf{101.15} \tag{2.44}$$

is the mle of θ. (Also note that because $E[Y] = \theta \implies E[\bar{Y}] = \theta$, $\hat{\theta}$ is an unbiased estimator of θ.)

```
> mean(y)
[1] 101.15
> abline(v = mean(y), lwd=3, lty=2)
> z=dgamma(y, shape=1, scale=mean(y))
> lines(z~y,lty=2)
```



(c) The sample median of the data is **55.5**

```
> median(y)
[1] 55.5
```

The median of $Y \sim \Gamma(1,\theta) \equiv \text{Exp}(\theta)$ is the value of y' at which

$$0.5 = \int_0^{y'} \frac{1}{\theta} e^{-y/\theta}\, dy = 1 - e^{-y'/\theta} \implies y' = \theta \ln 2, \tag{2.45}$$

which means that the median of $Y \sim \Gamma(1,\theta) \equiv \text{Exp}(1,\theta)$ is the half-life, $\theta \ln 2$. Since the sample median is just θ multiplied by $\ln 2$, the sample median also estimates the parameter θ.

(d) From part (a), we know that $\hat{\theta} = \bar{Y}$, the sample mean, is the mle of θ, the population mean. From part (c), we have shown that the median of $Y \sim \Gamma(1,\theta)$ is simply $\theta \ln 2$. By simple inspection we see that $\hat{\theta} \ln 2 = \hat{Y} \ln 2$ is the (*unbiased*) mle of $\theta \ln 2$, the median of Y. □

4.1.3 Suppose the number of customers X that enter a store between the hours 9:00 a.m. and 10:00 a.m. follows a Poisson distribution with parameter θ. Suppose a random sample of the number of customers that enter the store between 9:00 a.m. and 10:00 a.m. for 10 days results in the values

9 7 9 15 10 13 11 7 2 12

1. Determine the maximum likelihood estimate of θ. Show that it is an unbiased estimator.
2. Based on these data, obtain the realization of your estimator in part (a). Explain the meaning of this estimate in terms of the number of customers.

Solution:

1. Let $X \sim \text{Poi}(\theta)$ be given, then the pmf of X is given by

$$p(x) = \begin{cases} \frac{\theta^x e^{-\theta}}{x!}, & x \in \mathbb{N} \\ 0, & \text{else} \end{cases} . \tag{2.46}$$

Assuming the X_i's $\sim \text{Poi}(\theta)$ are iid, where $i = 1, \ldots, n$, then the logarithm of the likelihood function is

$$\begin{aligned} l(\theta) &= \log\left(\prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{x_i!}\right) \\ &= \log\left(e^{-n\theta}\theta^{\sum_{i=1}^{n} x_i} \prod_{i=1}^{n} \frac{1}{x_i!}\right) \\ &= -n\theta + \left(\sum_{i=1}^{n} x_i\right)\log\theta - \sum_{i=1}^{n} \log x_i!. \end{aligned} \tag{2.47}$$

Setting $\partial_\theta l(\theta) = 0$, we solve for θ:

$$\partial_\theta l(\theta) = -n + \frac{1}{\theta}\sum_{i=1}^{n} x_i = 0 \iff \theta = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \tag{2.48}$$

By inspection, $\partial_{\theta\theta} l(\theta) < 0 \forall \theta \in \mathbb{R}^+$, and so the statistic

$$\hat{\theta} = \bar{Y} \tag{2.49}$$

is the mle of θ. Further, it is an unbiased estimator of θ simply because

$$E[Y] = \theta \implies E[\bar{Y}] = \theta. \tag{2.50}$$

□

2. Part (a) says the sample means is the mle of θ. The means of the given sample is **9.5**.

```
> mean(c(9 ,7, 9, 15, 10, 13, 11, 7, 2, 12))
[1] 9.5
```

This says that on average, 9.5, or about 9-10 customers enter the store between the hours 9:00 a.m. and 10:00 a.m.. □

4.1.8 Recall that for the parameter $\eta = g(\theta)$, the mle of η is $g(\hat{\theta})$, where $\hat{\theta}$ is the mle of θ. Assuming that the data in Example 4.1.6 were drawn from a Poisson distribution with mean λ, obtain the mle of λ and then use it to obtain the mle of the pmf. Compare the mle of the pmf to the nonparametric estimate. Note: For the domain value 6, obtain the mle of $P(X \geq 6)$.

Solution: Based on the previous problem, the mle of λ is the sample means, which has the value **2.13**.

```
> mean(c(2,1,1,1,1,5,1,1,3,0,2,1,1,3,4,2,1,2,2,6,5,2,3,2,4,1,3,1,3,0))
[1] 2.133333
```

Because the sample means $\bar{x}$ is the mle of λ, and the pmf is given by

$$p(x) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!}, & x \in \mathbb{N} \\ 0, & \text{else} \end{cases} , \tag{2.51}$$

the mle of the pmf is given by

$$\tilde{p}(x) = \begin{cases} \frac{\bar{x}^x e^{-\bar{x}}}{x!}, & x \in \mathbb{N} \\ 0, & \text{else} \end{cases} . \tag{2.52}$$

Next, we compare the mle of the pmf to the nonparametric estimate:

j	0	1	2	3	4	5	≥ 6
$\hat{p}(j)$	0.067	0.367	0.233	0.167	0.067	0.067	0.033
$\tilde{p}(j)$	0.118	0.253	0.270	0.192	0.102	0.044	0.022

Mathematica code for $P(j \geq 6)$ for $\tilde{p}(j)$:

```
P[x_] := (2.1333333)^x*E^(-2.1333333)/x!

N[Sum[P[y], {y, 6, Infinity}]]

0.0218705
```