From Casella and Berger, do the following problems from Chapter 5:

Sections 5.1 and 5.2: 1, 3, 5, 6, 8, and 11.
Sections 5.3 and 5.4: 12, 13, 16, 17, 24, and 27.
Section 5.5: 30, 35, 36, 42, and 44.

These are extra problems that I have given on past exams (in STAT 712 or in related courses). You do not have to turn these in.

5.1. *Gamma distribution fun.*
(a) Suppose that $X \sim \text{gamma}(\alpha, \beta)$. Show that

$$Y = \frac{2X}{\beta} \sim \chi^2_{2\alpha}.$$

(b) Suppose $X_1 \sim \text{gamma}(\alpha_1, \beta_1)$, $X_2 \sim \text{gamma}(\alpha_2, \beta_2)$, and X_1 and X_2 are independent. Find a function of X_1 and X_2 that has an F distribution. What are the degrees of freedom?
(c) Suppose $X_1, X_2, ..., X_n$ are iid exponential(β_1) and $Y_1, Y_2, ..., Y_m$ are iid exponential(β_2). Suppose the samples are independent. Under the assumption that $\beta_1 = \beta_2$, find a function of $\overline{X}$ and $\overline{Y}$ that has an F distribution. What are the degrees of freedom?

5.2. *Normal distribution fun.* Suppose that $X_1, X_2, ..., X_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution. Let $\overline{X}$ and S^2 denote the sample mean and sample variance, respectively.
(a) Find the distribution of

$$Q = \frac{(\overline{X} - \mu)^2}{\sigma^2/n} + \frac{(n-1)S^2}{\sigma^2}.$$

(b) Formulate an approach to determine the value of c that satisfies

$$P\left(-c < \frac{S}{\overline{X}} < c\right) = 1 - \alpha,$$

for $\alpha \in (0, 1)$. The statistic $T(\mathbf{X}) = S/\overline{X}$ is called the *coefficient of variation* (it is a measure of variation, relative to the mean of the distribution).
(c) Suppose that X_{n+1} is an additional observation, which is distributed as $\mathcal{N}(\mu, \sigma^2)$ and is independent of $X_1, X_2, ..., X_n$ (i.e., X_{n+1} is a "new" observation). Prove that

$$T = \frac{X_{n+1} - \overline{X}}{S\sqrt{1 + 1/n}} \sim t(n-1).$$

5.3. *Noncentral F distribution ~~torture~~ fun.* The univariate random variable W is said to have a noncentral F distribution with degrees of freedom $n_1 > 0$ and $n_2 > 0$ and noncentrality parameter $\lambda > 0$ if it has the pdf

$$f_W(w|n_1, n_2, \lambda) = \sum_{j=0}^{\infty} \left(\frac{e^{-\lambda}\lambda^j}{j!}\right) \frac{\Gamma(\frac{n_1+2j+n_2}{2}) \left(\frac{n_1+2j}{n_2}\right)^{(n_1+2j)/2} w^{(n_1+2j-2)/2}}{\Gamma(\frac{n_1+2j}{2})\Gamma(\frac{n_2}{2}) \left(1 + \frac{n_1 w}{n_2}\right)^{(n_1+2j+n_2)/2}} I(w > 0).$$

We write $W \sim F_{n_1,n_2}(\lambda)$. When $\lambda = 0$, the noncentral F distribution reduces to the central F distribution.

(a) If U_1 and U_2 are independent random variables with $U_1 \sim \chi^2_{n_1}(\lambda)$ and $U_2 \sim \chi^2_{n_2}$, show that

$$W = \frac{U_1/n_1}{U_2/n_2} \sim F_{n_1,n_2}(\lambda).$$

Hint: Introduce $U = U_1 + U_2$ as a dummy variable and perform a bivariate transformation.
(b) If $W \sim F_{n_1,n_2}(\lambda)$, show that

$$E(W) = \frac{n_2}{n_2 - 2}\left(1 + \frac{2\lambda}{n_1}\right)$$

and

$$\text{var}(W) = \frac{2n_2^2}{n_1^2(n_2-2)}\left\{\frac{(n_1+2\lambda)^2}{(n_2-2)(n_2-4)} + \frac{n_1+4\lambda}{n_2-4}\right\}.$$

Note that $E(W)$ exists only when $n_2 > 2$ and var(W) exists only when $n_2 > 4$. The moment generating function for the noncentral F distribution does not exist in closed form.
(c) Suppose that $W \sim F_{n_1,n_2}(\lambda)$. For fixed n_1, n_2, and $c > 0$, show the quantity $P(W > c)$ is a strictly increasing function of λ. That is, if $W_1 \sim F_{n_1,n_2}(\lambda_1)$ and $W_2 \sim F_{n_1,n_2}(\lambda_2)$, where $\lambda_2 > \lambda_1$, then $P(W_2 > c) > P(W_1 > c)$.

Remark: That the noncentral F distribution tends to be larger than the central F distribution is the basis for many tests used in linear models. Typically, test statistics will have a central F distribution when H_0 is true and a noncentral F distribution when H_0 is not true. Because the noncentral F distribution tends to be larger, large values of a test statistic are consistent with H_1. Thus, the form of an appropriate rejection region is to reject H_0 for large values of the test statistic. The power is simply the probability of rejection region (defined under H_0) when the probability distribution is noncentral F. Noncentral F distributions are available in most software packages (e.g., R, etc.).

5.4. Suppose that $X_1, X_2, ..., X_n$ is an iid sample from

$$f_X(x|\mu) = e^{-(x-\mu)}I(x > \mu).$$

(a) Show that $\{f_X(x|\mu); -\infty < \mu < \infty\}$ is a location family.
(b) Show that the probability density function of the minimum order statistic $X_{(1)}$ is

$$f_{X_{(1)}}(x|\mu) = ne^{-n(x-\mu)}I(x > \mu).$$

(c) Show that $X_{(1)}$ converges in probability to μ.

5.5. Suppose $X_1, X_2, ..., X_n$ is an iid sample from $f_X(x) = e^{-x}I(x > 0)$.
(a) Show that

$$\begin{aligned} Z_1 &= nX_{(1)} \\ Z_2 &= (n-1)(X_{(2)} - X_{(1)}) \\ Z_3 &= (n-2)(X_{(3)} - X_{(2)}) \\ &\vdots \\ Z_n &= X_{(n)} - X_{(n-1)} \end{aligned}$$

are mutually independent and that each Z_i follows an exponential distribution marginally.
(b) Demonstrate that all linear combinations of $X_{(1)}, X_{(2)}, ..., X_{(n)}$ can be expressed as a linear combination of independent random variables.

5.6. Suppose that $Z_1, Z_2, ..., Z_6$ is an iid sample from a $\mathcal{N}(0,1)$ distribution. Suppose that $Z_7 \sim \mathcal{N}(0,1)$ and that Z_7 is independent of $Z_1, Z_2, ..., Z_6$. Find the distribution of

(a) $\overline{Z} = \frac{1}{6}\sum_{i=1}^{6} Z_i$

(b) $T = \sum_{i=1}^{6}(Z_i - \overline{Z})^2$

(c) $U = \sqrt{3}Z_7/\sqrt{Z_1^2 + Z_2^2 + Z_3^2}$

(d) $V = (Z_1^2 + Z_2^2 + Z_3^2)/(Z_4^2 + Z_5^2 + Z_6^2)$.

5.7. Let $X \sim F_{r,s}$; i.e., the probability density function (pdf) of X is

$$f_X(x) = \frac{\Gamma(\frac{r+s}{2})\left(\frac{r}{s}\right)^{r/2} x^{(r-2)/2}}{\Gamma(\frac{r}{2})\Gamma(\frac{s}{2})\left(1 + \frac{rx}{s}\right)^{(r+s)/2}} I(x > 0).$$

(a) Derive the pdf of $Y = 1/X$. Do not just state the answer.
(b) Carefully argue that $X \xrightarrow{p} 1$, as $\min\{r, s\} \to \infty$.

5.8. I have four statistics T_1, T_2, T_3, and T_4. I have determined that

- $T_1 \sim \mathcal{N}(0,1)$
- $T_2 \sim \mathcal{N}(-3,4)$
- $T_3 \sim \chi_3^2$
- $T_4 \sim \chi_5^2$
- T_1, T_2, T_3, and T_4 are mutually independent.

(a) What is the distribution of $3T_1 - 2T_2$?
(b) Find a statistic that has a t_8 distribution.
(c) Find a statistic that has an $F_{4,6}$ distribution.

5.9. Suppose that X and Y are independent random variables and let $Z = X + Y$. Use the method of convolution to derive the pdf of Z if
(a) X and Y are exponential(β)
(b) X and Y are $\mathcal{N}(0,1)$.
Note: In each case, it is easy to derive the distribution using mgfs, but this exercise is intended for you to get practice in applying the convolution technique.

5.10. Suppose X_1, X_2, X_3 are iid beta$(2,1)$.
(a) Derive the probability density function (pdf) of $X_{(1)}$.

(b) Let m denote the median of the beta$(2,1)$ distribution. Find $P(X_{(1)} > m)$.
(c) Calculate $\text{cov}(X_{(2)}, X_{(3)})$.

5.11. Suppose that $X_1, X_2, ..., X_n$ are iid exponential(1). Define the statistics

$$\begin{aligned} U_n &= 2(X_1 + X_2 + \cdots + X_n) \\ V_n &= \sqrt{n}(\overline{X} - 1). \end{aligned}$$

(a) Derive the moment generating function (mgf) of U_n. What is the distribution of U_n?
(b) Show that the mgf of V_n, for $t < \sqrt{n}$, is

$$M_{V_n}(t) = \{e^{t/\sqrt{n}} - (t/\sqrt{n})e^{t/\sqrt{n}}\}^{-n}.$$

(c) For $t < \sqrt{n}$, find $\lim_{n\to\infty} M_{V_n}(t)$.

5.12. Suppose that $X_1, X_2, ..., X_n$ are iid Poisson(θ) and let $T = \sum_{i=1}^n X_i$ denote the sample sum.
(a) Find the conditional distribution of X_1 given $T = t$.
(b) Show that T has pdf in the exponential family.
(c) Show that $\overline{X}$ has pdf in the exponential family.

5.13. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, where $-\infty < \mu < \infty$ and $\sigma^2 > 0$.
(a) Consider the statistic cS^2, where c is a constant and S^2 is the usual sample variance (denominator $= n-1$). Find the value of c that minimizes

$$\text{MSE}(cS^2) = E[(cS^2 - \sigma^2)^2] = \text{var}(cS^2) + [E(cS^2 - \sigma^2)]^2.$$

(b) Consider the normal subfamily where $\sigma^2 = \mu^2$, where $\mu > 0$. Let S denote the sample standard deviation. Find a linear combination $c_1\overline{X} + c_2 S$ whose expectation is equal to μ. Find the variance of this linear combination.

5.14. Suppose that $Z_0, Z_1, Z_2, ...,$ is a sequence of iid standard normal random variables, and define, for $i = 1, 2, ...,$

$$X_i = Z_0 + Z_i,$$

Show that $\overline{X}_n \xrightarrow{p} Z_0$, as $n \to \infty$.

5.15. *Monte Carlo approximation.* Suppose that you want to compute

$$I = E\left[Y \sin\left(2\pi \cos Y\right)\right],$$

where $Y \sim \mathcal{U}(-\pi/2, \pi/2)$. Approximate this expectation by computing

$$\widehat{I}(Y_1, Y_2, ..., Y_n) = \frac{1}{n}\sum_{i=1}^n Y_i \sin\left(2\pi \cos Y_i\right),$$

for large n, where $Y_1, Y_2, ..., Y_n$ is an iid $\mathcal{U}(-\pi/2, \pi/2)$ sample. What is your approximation of I? How good do you think the approximation is? Could you use the same method to approximate $E(\tan Y)$? Why or why not?

5.16. Suppose that the bivariate vector $\mathbf{Y} = (Y_1, Y_2)'$ has the following density:

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} e^{-y_1}, & 0 < y_2 < y_1 < \infty \\ 0, & \text{otherwise.} \end{cases}$$

We observe n iid copies of $\mathbf{Y}$, say, $\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_n$, where $\mathbf{Y}_j = (Y_{1j}, Y_{2j})'$, for $j = 1, 2, ..., n$. Define

$$\overline{\mathbf{Y}}_n = (\overline{Y}_{1+}, \overline{Y}_{2+})' \equiv \left(\frac{1}{n} \sum_{j=1}^{n} Y_{1j}, \frac{1}{n} \sum_{j=1}^{n} Y_{2j} \right)'$$

and $\boldsymbol{\mu} = (\mu_1, \mu_2)'$, where $\mu_i = E(Y_{i1})$, for $i = 1, 2$.
(a) Find a centered and scaled version of $\overline{\mathbf{Y}}_n$ that converges in distribution to $\mathbf{Y} \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma})$, as $n \to \infty$, and compute $\boldsymbol{\Sigma}$.
(b) When $n = 30$, approximate $P(\overline{Y}_{1+} - \overline{Y}_{2+} > 0.5)$.

5.17. Suppose that $\mathbf{X} = (X_1, X_2, X_3)' \sim \text{mult}(n, \mathbf{p} = \frac{1}{3}\mathbf{1})$, where $\mathbf{1} = (1, 1, 1)'$.
(a) Show that

$$\mathbf{Y}_n = \frac{\mathbf{X} - n\mathbf{p}}{\sqrt{n}} \xrightarrow{d} \mathcal{N}_3(\mathbf{0}, \boldsymbol{\Sigma}),$$

as $n \to \infty$, and compute $\boldsymbol{\Sigma}$.
(b) For $n = 30$, first approximate $P(|X_1 - X_2| \leq 1)$ using the result in (a). Then, compute this probability exactly using the multinomial pmf. How close are the two answers?

5.18. Suppose that $X_1, X_2, ..., X_n$ is an iid gamma$(\alpha, 1)$ sample, and let $\overline{X}_n$ denote the sample mean. Define

$$Z_n = \frac{\sqrt{n}\,(\overline{X}_n - \alpha)}{\sqrt{\overline{X}_n}}.$$

(a) Prove that $Z_n \xrightarrow{d} Z$, where $Z \sim \mathcal{N}(0, 1)$.
(b) Argue that

$$\left(\overline{X}_n - 1.96\sqrt{\frac{\overline{X}_n}{n}},\ \overline{X}_n + 1.96\sqrt{\frac{\overline{X}_n}{n}} \right)$$

serves as a large-sample 95 percent confidence interval for α.
(c) Show that Z_n^2 converges in distribution, and find the distribution.

5.19. Suppose that $X_1, X_2, ..., X_n$ is an iid exponential(θ) sample, where $E(X_1) = \theta$ is unknown, and define $Y_i = X_i^2$, for $i = 1, 2, ..., n$.
(a) Use the CLT to derive large-sample distribution of a properly centered and scaled version of $(\overline{X}, \overline{Y})'$.

(b) Find a consistent estimator of the covariance matrix in part (a). For the most part, "consistency" means "convergence in probability."

5.20. Suppose that $X_1, X_2, ..., X_n$ is an iid sample with mean $\theta > 0$. Under a Poisson model, $E(X_1) = \text{var}(X_1) = \theta$. Therefore, a standard test of the hypothesis that

$$H_0 : X_1, X_2, ..., X_n \text{ is an iid Poisson sample}$$

uses $T = T(\mathbf{X}) = S_n^2/\overline{X}$ and rejects H_0 if T is too large. This test is good against alternatives whose variance is greater than the mean, such as the negative binomial distribution or any other mixture of Poisson distributions.
(a) Derive the large-sample distribution of T, properly centered and scaled, under the Poisson assumption.
(b) Test the Poisson model assumption with the following data, which denote the number of airborne mold spores collected from specimens of air in my office (just kidding?).

4	8	10	6	14	7	5	11	5	17

Hint: To perform the test in (b), you'll want to find a suitably normalized version of T that converges to a standard normal distribution as n gets large. Remember that your statistic can not depend on θ.

5.21. Suppose that $X_1 \sim b(n, p_1)$, $X_2 \sim b(n, p_2)$, and that X_1 and X_2 are independent. Define the population odds ratio by

$$\theta = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

and the sample odds ratio by

$$\widehat{\theta} = \frac{\widehat{p}_1/(1-\widehat{p}_1)}{\widehat{p}_2/(1-\widehat{p}_2)}.$$

(a) Find the large-sample distribution of $\widehat{\theta}$, properly centered and scaled.
(b) Find the large-sample distribution of $\ln \widehat{\theta}$, properly centered and scaled.
(c) Use simulation to create qq plots of $\widehat{\theta}$ and $\ln \widehat{\theta}$ for different sample sizes n and different values of p_1 and p_2 (don't go overboard). Which statistic, $\widehat{\theta}$ or $\ln \widehat{\theta}$, seems to "converge to normality" faster?

5.22. Suppose that $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ is an iid sample (with appropriate moments existing) from a distribution with correlation ρ.
(a) Show that the sample correlation coefficient r can be written as a function of (uncentered) sample moments

$$\overline{X},\ \overline{Y},\ \frac{1}{n}\sum_{i=1}^{n} X_i^2,\ \frac{1}{n}\sum_{i=1}^{n} Y_i^2,\ \text{and}\ \frac{1}{n}\sum_{i=1}^{n} X_i Y_i.$$

(b) Explain how to get the asymptotic distribution of r using the multivariate Delta Method. Give as many details as possible, but you don't have to carry out the matrix multiplication.

(c) When the population distribution is bivariate normal,

$$\sqrt{n}(r-\rho) \xrightarrow{d} \mathcal{N}[0,(1-\rho^2)^2].$$

Derive a transformation g so that the asymptotic variance of $g(r)$ is free of ρ. This is called a *variance-stabilizing transformation.*

5.23. Let Y have an F distribution with r numerator and s denominator degrees of freedom; i.e., $Y \sim F(r,s)$.
(a) Show that $rY \xrightarrow{d} \chi^2(r)$, as $s \to \infty$.
(b) For specificity, assume that $Y \sim F(3,96)$. Give an applied situation where this distribution would arise (e.g., think of one-way analysis of variance). Explain how you could use the result from part (a) to approximate $P(Y \geq 1.25)$. How good is this approximation?

5.24. Suppose that $X_1, X_2, ..., X_n$ is an iid sample from a Poisson distribution with mean $\theta > 0$.
(a) Why does $\overline{X}_n \xrightarrow{p} \theta$, as $n \to \infty$? Just state the name of the result.
(b) Consider the function $g(\theta) = (\theta+1)\exp(-\theta)$. Prove that, as $n \to \infty$,

$$\sqrt{n}\left\{g(\overline{X}_n) - g(\theta)\right\} \xrightarrow{d} \mathcal{N}\left\{0, \theta^3 \exp(-2\theta)\right\}.$$

(c) Find two statistics that converge in probability to $\theta^3 \exp(-2\theta)$.

5.25. Suppose that $X_1, X_2, ..., X_n$ is an iid sample from

$$f_X(x|\theta) = \frac{1}{2\theta} I(0 < x < 2\theta),$$

where $\theta > 0$.
(a) Find a function of $X_{(n)}$ that converges in probability to θ, as $n \to \infty$. Prove your claim.
(b) Find a function of $\overline{X}$ that converges in probability to θ, as $n \to \infty$. Prove your claim.
(c) Find a function of $\overline{X}$ that converges in distribution to a normal distribution as $n \to \infty$. Find the mean and variance of the limiting distribution.

5.26. Suppose that $X_1, X_2, ..., X_n$ is an iid sample from a $\mathcal{U}(0,1)$ distribution. Define the statistics

$$\begin{aligned}
\widehat{\theta}_A &= \frac{1}{n}\sum_{i=1}^{n} X_i \\
\widehat{\theta}_G &= \left(\prod_{i=1}^{n} X_i\right)^{1/n} \\
\widehat{\theta}_H &= \left(\frac{1}{n}\sum_{i=1}^{n} \frac{1}{X_i}\right)^{-1}.
\end{aligned}$$

(a) Which statistics converge in probability to a finite constant (as $n \to \infty$)? For those that do converge in probability, find the limiting constant. For those that do not, explain why.

(b) If possible, find a function of each statistic that converges to a standard normal distribution. Work with each statistic separately (do not combine them).

5.27. Suppose that $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(0, \sigma^2)$, where $\sigma^2 > 0$.
(a) Find a function of

$$T \equiv T(\mathbf{X}) = \sum_{i=1}^{n} X_i^2$$

that converges in distribution to a normal distribution as $n \to \infty$. State the mean and variance of your limiting normal distribution.
(b) Find a function of T whose large-sample variance is free of σ^2.